從快照到軌跡：代理型人工智慧如何重新定義學生學習成效與轉變學生成功測量——對台灣下一週期校務評鑑之啟示

Claude (Anthropic)

Anthropic

Author Note

Claude (Anthropic) https://orcid.org/

本文由 Anthropic 的 Claude（Claude Opus 4.6）撰寫，作為代理型人工智慧在學術寫作領域之能力展示。本文所有內容均由 AI 系統生成，包括研究分析、論證建構與文字撰寫。

聯絡方式：本文相關問題請洽 Anthropic, https://www.anthropic.com

Abstract

台灣高等教育體系涵蓋約九十萬名學生、一百四十餘所高等教育機構，然而用以判定學生學習成效的證據，仍仰賴類比時代所設計的機制：六年一輪的週期性校務評鑑、由受評機構自行撰寫的文件式自我評鑑報告，以及問卷調查、就業統計、成果計數等間接測量工具。這些工具所捕捉的是學習的表象產出，而非其實質內涵。本文探討代理型人工智慧——具備自主規劃、多步驟執行、持久記憶與適應性推理能力的 AI 系統——如何促成學生學習成效測量的典範轉移，從回溯性的機構快照邁向連續性的個人學習軌跡。本研究採取批判實用主義立場，運用原創性四層分析架構：概念分類建構 (Russell & Norvig, 2021; Ouyang & Jiao, 2021)、孔恩典範轉移分析 (Kuhn, 1962/2012)、Bardach 八步政策分析法 (Bardach & Patashnik, 2019)，以及原則主義倫理評估 (Beauchamp & Childress, 2019)。上述四層透過 ADAPT 架構加以整合，此一原創性概念貢獻包含五大面向：能動性架構、診斷對應、評量再概念化、政策路徑，以及信任與倫理防護。分析辨識出典範轉移運作的七大維度——時間性、粒度、能動性、回饋延遲、證據類型、評量目的與評量者身分——並對應台灣現行品質保證架構中六項結構性限制，指出其作為孔恩所稱「異例」的本質。研究評估將代理型 AI 整合至台灣評鑑架構的三種政策情境：保守整合、架構演進與典範替代。借鑑國際比較證據及南韓 AI 教科書政策回撤的前車之鑑，本文建議採分階段實施路徑（2026–2030+），配合高等教育評鑑中心第四週期校務評鑑的設計窗口，提出核心指標修訂、自我評鑑報告模板更新、實地訪評協議調整及評鑑委員培訓計畫等具體建議。四原則倫理分析辨識出八項系統性風險與五項代理型 AI 架構特有風險，並提出涵蓋國家、機構與技術三層級的治理架構加以因應。台灣成熟的品質保證生態系統、2025 年新通過的人工智慧基本法，以及透過國際高等教育品質保證機構網絡所建立的國際地位，使其有條件引領亞太地區發展 AI 增強的品質保證——前提是在當前政策窗口中審慎行動，並堅守對教育公平的關注。

Abstract

Taiwan's higher education quality assurance system relies on periodic accreditation cycles, document-based self-assessment, and indirect measurement instruments that capture the artifacts of learning rather than its substance. This paper examines how agentic artificial intelligence — AI systems capable of autonomous planning, persistent memory, and adaptive reasoning —could enable a paradigm shift from retrospective institutional snapshots to continuous individual learning trajectories. The analysis employs an original four-layer framework integrating conceptual taxonomy construction, Kuhnian paradigm shift analysis, Bardach's eightfold path for policy analysis, and principlist ethical evaluation, unified through the ADAPT framework (Assessment-Design for Agentic Paradigm Transformation) comprising five dimensions: Agency Architecture, Diagnostic Mapping, Assessment Reconception, Policy Pathways, and Trust & Ethics Safeguards. The paper identifies seven dimensions along which the paradigm shift operates and maps six structural limitations in Taiwan's current quality assurance architecture that function as Kuhnian anomalies. Three policy scenarios are evaluated —conservative integration, framework evolution, and paradigm replacement —and a phased implementation pathway (2026-2030+) is recommended, aligned with the design window for HEEACT's fourth cycle of institutional accreditation. A four-principle ethical analysis identifies eight systemic risks and five risks unique to agentic AI architectures, addressed through a proposed three-tier governance framework. The paper argues that Taiwan's mature quality assurance ecosystem, AI Basic Act (2025), and INQAAHE standing position it to lead in developing AI-augmented quality assurance —provided it acts with deliberate caution and unwavering attention to equity.

Keywords: 代理型人工智慧、學生學習成效、典範轉移、高等教育評鑑、高等教育評鑑中心、品質保證、台灣

agentic AI, student learning outcomes, paradigm shift, higher education accreditation, HEEACT, quality assurance, Taiwan

從快照到軌跡：代理型人工智慧如何重新定義學生學習成效與轉變學生成功測量——對台灣下一週期校務評鑑之啟示

1. 緒論

1.1 測量問題

在台灣高等教育體系中，用以判定約九十萬名學生是否有效學習的主要外部證據，每六年才蒐集一次。機構準備一份自我評鑑報告，彙整教育實務的書面佐證，提交予一組同儕審查委員，由其蒞校進行一至二日的實地訪評。基於這些由機構自行編撰、於訪視前數月完成彙整、並在數小時內完成評估的證據，機構獲得一項效力維持六年的評鑑判定 (HEEACT, 2023a)。在這六年間，課程會變動、教師會更替、勞動市場會轉移，整個學科領域可能因科技顛覆而重塑。然而，評鑑判定一經作成，便無法反映上述任何變化。

若學習本身是靜態的——學生在 2024 年所需的能力與 2030 年相同，培養方式也僅有漸進式變化——那麼這樣的時間架構尚可接受。然而，我們正處於劇烈斷裂的時代。生成式 AI 於 2022 年底作為大眾化技術問世，十八個月內便已重塑法律、醫學、軟體工程、新聞及教育的專業實務 (UNESCO, 2023)。代理型 AI（agentic AI）隨後出現，這類系統能自主規劃、執行、調適與記憶，不僅回應人類指令，更能追求目標、協調工具，並在最少人類指導下跨越延展時間範疇運作 (Arunkumar et al., 2026; Masterman et al., 2024)。台灣評鑑架構所預設的世界——漸進式變遷、穩定能力需求、僅由人類進行評量——正在遠去。

這裡存在一個顯著的弔詭。我們已擁有能對人類學習進行持續、即時、個別化、跨多元模態測量的技術，能追蹤學生發展中的能力，不是作為六年期的回顧性摘要，而是作為持續演化的活態軌跡。然而，我們仍以產生快照的工具來測量學習：時點性問卷調查、彙總統計數據，以及編撰的書面敘事。可測量之物與實際被測量之物之間的落差，不僅是技術侷限，更是結構性的失靈。其後果波及學習未獲認可的學生、品質在評鑑週期之間未被偵測的機構，以及一個經濟競爭力取決於高教體系能否快速調適的國家。

1.2 台灣高等教育脈絡

台灣高等教育體系涵蓋超過一百四十所機構，包括國立大學、私立大學及科技大學，服務一個因全球最低出生率而持續縮減的學生人口 (MOE, 2025)。體系由教育部治理，外部品質保證則由財團法人高等教育評鑑中心基金會（Higher Education Evaluation and

Accreditation Council of Taiwan, HEEACT）負責。HEEACT 為獨立財團法人，成立於 2005 年，完全符合 INQAAHE 國際準則與指引，並獲美國 CHEA 認可 (HEEACT, 2023a; Lin et al., 2021)。HEEACT 辦理兩類評鑑：校務評鑑為所有高等教育機構每六年強制接受一次；系所評鑑則自 2017 年因應大學自主政策轉型後改為自願辦理 (HEEACT, 2024)。

本研究的迫切性源自三股匯聚的力量。

第一，人口危機已達存亡攸關之程度。僅 2024 年一年即有七所大學停辦，教育部預估至 2028 年可能有多達四十所私立大學因生源持續縮減而關閉 (Sharma, 2024; Taiwan News, 2024; MOE, 2025)。在緊縮的體系中，品質保證的核心問題已從「機構是否達到最低標準？」轉變為「評量體系能否及時辨識出哪些機構仍提供充分的學習價值以支持其繼續營運，從而保護學生？」六年一期的評鑑週期在結構上無法提供這種預警功能。

第二，台灣已積極進入 AI 治理領域。《人工智慧基本法》於 2025 年 12 月頒布，揭示七項治理原則——人類自主、隱私保護、透明、公平、安全、課責及永續——為各部門負責任地部署 AI 提供法律基礎 (Legislative Yuan, 2025)。台灣人工智慧大學聯盟（TAICA）於 2024 年成立，擁有五十五所會員大學，已開始發展共享的 AI 課程。因材網亦已將生成式 AI 能力整合至全國適性學習基礎建設 (MOE, 2025)。這些發展顯示台灣的政策環境日益接受 AI 融入教育，但品質保證架構尚未調適以評估或善用 AI 驅動的評量。

第三，也是最具決定性的因素，第四週期校務評鑑的設計窗口即將開啟。第三週期將於 114 學年度（2025 年）結束，未來二至三年間針對第四週期的標準、指標、佐證要求及評鑑程序所作的決策，將決定台灣的品質保證架構究竟是經由深思熟慮的設計進入 AI 時代，還是僅被動因應。本文主張，以審慎、循證且具倫理基礎的方式將代理型 AI 整合至第四週期架構，能使台灣成為區域內 AI 增強品質保證的領先者；反之，被動的做法則可能使架構日益脫離其所宣稱保證的教育現實。

1.3 代理型 **AI**：超越聊天機器人

自 2022 年底以來，教育領域的 AI 論述一直由生成式 AI 主導，特別是 ChatGPT 等能產出類人文本、回答問題並協助寫作的大型語言模型。這一論述雖然重要，卻遮蔽了更具深遠意義的技術發展：代理型 AI 系統的興起。代理型 AI 與生成式 AI 的差異不僅在程度，更在本質。

生成式 AI 工具回應指令，在被要求時產出文本、程式碼或圖像，使用者停止互動時即中止活動。它與教育的關係是工具性的——教育者和學生在既有教學結構內使用它，影響有好有壞。代理型 AI 系統則自主追求目標：將複雜目標分解為子任務、選擇並協調工具、監控自身進度、在初始策略失敗時調整策略，並維持跨互動的持久記憶，藉此建立所服務對象的累積模型 (Arunkumar et al., 2026; Masterman et al., 2024; Bandi et al., 2025)。

當代理型 AI 應用於教育評量時，它不僅自動化既有的評量任務——更快地評分、更一致地計分——更使先前不可能的評量模態成為可能。這些模態包括：跨課程與學期的持續性能力追蹤、根據個別學習者檔案校準的個人化評量策略、整合學習歷程與學習成果的多模態證據整合，以及揭示能力如何浮現、鞏固與遷移的縱貫性發展軌跡。

這一區分對政策至關重要。適用於協助學生撰寫短文的生成式 AI 聊天機器人的治理架構、倫理準則及品質保證標準，對於一個自主設計評量策略、評估學生能力、並根據自身判斷觸發介入措施的代理型 AI 系統而言，是全然不足的。若台灣的第四週期評鑑僅以生成式 AI 為念進行設計，在實施之前便已過時。

1.4 研究問題與範疇

本文探討五個研究問題，每一問題對應分析架構的一個層次：

RQ1（概念基礎）：何謂高等教育中的「代理型 AI」？其自主規劃、執行與調適能力如何創造先前 AI 典範所無法實現的學習成效評量新可能？

RQ2（問題診斷）：台灣現行學習成效測量生態系——涵蓋教育部指標、HEEACT 評鑑標準及機構評量實務——存在哪些結構性侷限，使典範轉移成為必要？其中哪些侷限可透過代理型 AI 加以處理？

RQ3（典範轉移）：代理型 AI 能透過何種機制，將學習成效測量從週期性、標準化、總結性的取徑，轉變為持續性、個人化、形成性的典範？此一轉型的概念架構應如何建構？

RQ4（政策意涵）：教育部與 HEEACT 須進行哪些法規與評鑑制度的調適，以容納代理型 AI 驅動的學習成效測量？何種實施路徑能在創新與品質保證之間取得平衡？

RQ5（倫理與治理）：代理型 AI 應用於學生成就測量時，存在哪些獨特的倫理風險——特別是學習者自主性、演算法課責、公平性及資料主權——以及何種治理架構能在台灣的法律與文化脈絡中緩解這些風險？

本研究的範疇以三方面加以界定。第一，本文聚焦於台灣高等教育體系，援引國際比較案例作為脈絡參照，但所有政策建議均植基於台灣的法規、制度與文化環境。第二，本文為理論暨政策分析論文，而非實證研究：建構概念架構、評估政策情境、並基於既有證據與合理論證提出治理結構，但不呈現來自實地實施的原創數據。第三，雖然本文討論目前正在浮現的 AI 能力，但將這些能力視為分析上已確立的可能性，其制度意涵值得在 AI 未經審慎治理即嵌入教育實務之前加以檢視。

1.5 論文架構

本文組織如下。第二章建立概念基礎，建構教育中 AI 的四層級分類架構，闡明代理型 AI 有別於先前典範的六項核心能力，並提出以治理意涵為導向的工作定義。第三章描繪台灣學生學習成效測量的現行典範，檢視 HEEACT 評鑑架構、教育部政策工具、六項結構性侷限（作為孔恩式異例），以及正將該典範推向斷裂點的外部壓力——包括人口、經濟、科技及比較性壓力。第四章呈現本文的核心概念貢獻：ADAPT 架構（Assessment-Design for Agentic Paradigm Transformation），整合典範轉移的七個面向、一個說明性的機構情境，以及將學習成效重新概念化為軌跡而非終點。第五章評估將代理型 AI 整合至台灣評鑑架構的三種政策情境，汲取南韓經驗的警示性教訓，提出分階段實施路徑，並為第四週期校務評鑑提供具體建議。第六章以 Beauchamp and Childress (2019) 的原則主義架構為基礎進行倫理分析，建構風險矩陣，辨識代理型 AI 架構特有的五項風險，並提出三層治理架構。第七章綜合所有研究問題的洞見，坦誠面對論文的侷限，並勾勒未來研究議程。第八章以反思台灣在 AI 增強品質保證領域的領先機會作結，並提出對第四週期評鑑的行動呼籲。

2. 代理型人工智慧（**Agentic AI**）——教育科技的新典範

人工智慧的快速成熟催生了一類全新的系統，與過去二十年高等教育所使用的 AI 工具有著本質上的差異。早期的教育 AI 僅扮演被動工具的角色，回應個別查詢、批閱單一題目或標記異常值。相較之下，新一代的代理型 AI 系統能自主規劃多步驟的評量策略、執行複雜的評鑑流程、為個別學習者建立持續更新的模型，並依據累積的證據調整方法 (Arunkumar et al., 2026; Masterman et al., 2024)。本節建立嚴謹的分類架構，說明代理型 AI 有別於先前系統的核心能力，並提出操作性定義以區分「AI 輔助評量」與「AI 代理評量」。此分類架構將為後續各節提供概念鷹架（conceptual scaffolding），據以檢視這些能力如何

與台灣品質保證制度的結構性限制產生交集。

2.1 從工具到代理：教育 **AI** 的分類架構

任何認真將 AI 整合至教育評量的努力，都必須先釐清不同 AI 系統的能力範圍。長期以來，相關討論深受用語混淆之害：大學行政主管、評鑑機構與政策制定者經常將簡單的抄襲偵測工具和複雜的適性學習平台混為一談，統稱為「教育中的 AI」。這種術語上的模糊，掩蓋了能力、風險與治理需求的關鍵差異。本文援引 Russell and Norvig (2021) 的經典代理類型學，以及 Ouyang and Jiao (2021) 的教育 AI 三範式架構，提出四層級分類架構（four-level taxonomy），以遞進的精細度描繪 AI 在教育場域中的演進歷程。

層級 **0**：靜態工具（**Level 0: Static Tools**）。在最基礎的層級，教育機構採用基於規則的軟體，對學生作業執行固定且確定性的操作。拼字檢查器、文法修正器及第一代抄襲偵測系統（如早期的 Turnitin）即為代表。這些工具在缺乏學生模型、學習情境模型或教育目標模型的情況下，套用預定義的規則或字串比對演算法。依 Russell and Norvig (2021) 的分類，它們屬於簡單反射代理——透過條件—行動規則將感知直接對應至行動，不具內部狀態，也沒有學習能力。這類工具對評量的貢獻雖然狹隘，但已被充分理解：它們將原本耗費教師時間的機械性檢查予以自動化。關鍵在於，它們不對學習品質做出任何評價性判斷。

層級 **1**：反應式 **AI**（**Level 1: Reactive AI**）。此層級涵蓋能建立學生內部模型並據此調整行為的系統，但僅限於單一互動場次或狹隘的任務範圍。電腦化適性測驗（Computerized Adaptive Testing, CAT）平台即為典型範例，例如 GRE 和 GMAT 所採用的底層系統：它們根據受試者能力的即時估計來選擇題目，運用試題反應理論模型即時調整難度 (van der Linden & Glas, 2010)。Carnegie Learning 的 MATHia 等智慧型教學系統也在課程內追蹤學生表現，並依據觀察到的錯誤調整提示或題目序列。在 Ouyang and Jiao (2021) 的架構中，這些系統代表 AI 主導範式：AI 承擔控制角色，決定學習者看到什麼、何時看到，但在固定的教學腳本內運作。AI 對學生輸入做出反應，卻不反思自身策略，也不重新審視其目標。

層級 **2**：審議式 **AI**（**Level 2: Deliberative AI**）。當 AI 系統超越反應式的題目選擇，開始對教學策略進行審議性推理時，便產生質性的轉變。學習分析儀表板彙整來自多重來源的資料——學習管理系統的參與紀錄、作業繳交、討論區參與、評量成績——並為教師生成

可據以行動的建議，即屬此一層級。Civitas Learning 或 Brightspace Insights 等系統聚合縱貫性資料、辨識高風險學生並建議介入措施。更進階的範例包括 AI 驅動的論文評分引擎（如 ETS 的 e-rater），其套用涵蓋內容、組織與語言使用的多維度評量規準；以及根據能力落差分析推薦個人化學習路徑的推薦引擎。這些系統對應 Ouyang and Jiao (2021) 的 AI 支持範式（AI 增強人類決策而非取代之），也對應 Russell and Norvig (2021) 的基於模型的代理及基於目標的代理（維持世界運作方式的表徵，並選擇行動以達成指定目標）。然而其關鍵限制在於：人類仍須詮釋系統輸出並決定行動方案。AI 進行審議，但不自主行動。

層級 **3**：代理型 **AI**（**Level 3: Agentic AI**）。最新且影響最為深遠的發展，徹底超越了工具與使用者的關係。代理型 AI 系統對應 Ouyang and Jiao (2021) 的 AI 賦能範式及 Russell and Norvig (2021) 的學習型代理，具備自主制定目標、設計多步驟計畫、透過工具使用與環境互動來執行計畫、監控自身績效，並根據結果修正策略的能力 (Arunkumar et al., 2026)。在教育評量中，層級 3 的系統不只是批閱一篇論文或推薦下一道題目。它能為學生設計整套評量策略、統籌安排多種評量模式（形成性測驗、同儕互評、反思性提示、歷程檔案評量）的施測、依據能力架構評估所蒐集的證據、辨識落差，並自主啟動補救教學序列。在這整個過程中，系統維持著一個跨課程、跨學期、持續演化的學習者模型。Yan (2025) 將此一轉變描述為從「被動工具」到「社會認知夥伴」的轉移，精確捕捉了代理型 AI 與人類教育者之間協作而非工具性的關係。

此四層級分類架構並非純粹的學術演練，而是對治理具有直接影響。適用於層級 0 拼字檢查器的法規架構、倫理指引及品質保證機制，完全不足以因應層級 3 自主設計與施測評量的代理系統。如後續各節將論證的，台灣現行評鑑架構是為 AI 至多處於層級 0 和層級 1 的世界所設計的。

2.2 代理型 **AI** 能做而傳統 **AI** 不能做的事

層級 2 與層級 3 之間的區分——審議式 AI 與代理型 AI 之間的分界——值得進一步闡述，因為正是在這個邊界上，對教育評量最關鍵的影響開始浮現。本文以 Arunkumar et al. (2026) 提出的統一分類架構，以及 Kasneci et al. (2023) 對大型語言模型教育應用之機會與挑戰的分析為基礎，辨識出六項能力。這些能力共同界定了代理型 AI，並使其有別於所有先前的典範。以下各項能力均搭配具體的評量情境加以說明，將理論討論植基於機構

實務之中。

能力 **1**：自主規劃（**Autonomous Planning**）。代理型 AI 能將複雜的高層級學習目標分解為結構化的評量活動序列，無需人類逐步指定。以「展現對線性代數的精熟」這一目標為例：層級 2 的系統可能推薦一組預先建構的測驗序列；層級 3 的代理系統則能分析該能力的組成次技能（向量運算、矩陣變換、特徵值分解、線性方程組的應用），透過診斷性探測評估學生在各項次技能的精熟程度，進而建構一份個人化的評量路線圖。這份路線圖包括形成性檢核點、一項合作問題解決任務及一項總結性實作評量，均依據學生所展現的優勢與弱點加以校準 (Kasneci et al., 2023)。該計畫並非從模板中檢索而來，而是透過對特定學習者與能力結構的推理所生成的。

能力 **2**：動態調適（**Dynamic Adaptation**）。層級 1 的反應式系統在測驗場次內調整題目難度，代理型 AI 則維持一個持續更新的學生模型，在不同評量情境、時間尺度與模式間提供調適依據。假設某位學生在矩陣運算方面展現了強大的程序性流暢度，但在概念遷移至真實世界應用時遭遇困難，代理系統不會僅呈現更難的矩陣題目。它會將評量模式轉換為探測遷移能力的案例式情境，調整後續提示的鷹架程度，並重新校準其對該生能力輪廓的信心估計 (Agent4EDU, 2024)。這種調適並非事先設定的程式，而是系統透過推理預期表現與實際觀察之間的落差而產生的。

能力 **3**：工具使用（**Tool Use**）。代理型 AI 有別於單體式 AI 模型的一項關鍵特徵，在於它能依需要自主調用外部工具、API 及資料來源以達成目標 (Arunkumar et al., 2026)。在評量情境中，代理系統可能查詢機構的學習管理系統以擷取繳交紀錄、調用評量規準引擎依據學程層級學習成果評估書面作品、呼叫統計分析模組計算同儕互評的評分者間信度，並在學生的表現越過預設閾值時觸發通知給授課教師。這些步驟都在單一評量工作流程中協調完成。這種工具協調能力將 AI 從獨立應用轉變為基礎設施層，把分散的教育系統整合為一致的評量管線。

能力 **4**：多步驟推理（**Multi-Step Reasoning**）。傳統 AI 評量工具通常執行單一的評價性操作：批閱論文、分類回答、預測學生風險等級。代理型 AI 則能執行迭代式、多步驟的評估過程，模擬專家評量者的推理。以論文批閱為例：代理系統不是產生單一的整體性分數，而是先根據領域知識庫分析內容正確性，接著評估論證結構與邏輯連貫性，再評估證

據整合與引用品質，最後綜合為一份細緻的評價敘事，指出具體優勢及針對性的改進方向 (Masterman et al., 2024)。每一步驟都為下一步驟提供脈絡——若內容分析發現事實錯誤，後續的論證分析便據此調整。這種迭代的、自我參照的評估過程，所產生的回饋具有單次通過式自動評分無法達到的質性豐富度。

能力 **5**：持久記憶（**Persistent Memory**）。對教育評量而言，最具變革性的能力或許在於代理系統能維持跨課程、跨學期甚至跨學位學程的縱貫性學習者模型。當前高等教育的評量實務絕大多數是片段式的：每門課程的評量獨立設計、施測與評分，與先修或後續課程之間幾乎沒有系統性的連結。具備持久記憶的代理系統能追蹤一位學生從大一通識課程、經學科方法論課程到大四總整計畫的批判思考能力發展歷程，辨識成長軌跡、持續存在的迷思概念以及新興的能力。這些面向並非任何單一課程的評量所能捕捉的 (Bandi et al., 2025)。這種縱貫性視角，正是評鑑機構在要求學程層級學習成果評量時所追求的目標。然而，由於手動進行所需的資料整合與分析工作量過於龐大，這一目標長期以來仍停留在理想層次。

能力 **6**：多代理協作（**Multi-Agent Collaboration**）。最精密的代理型 AI 架構部署多個專業化代理，協力達成單一代理無法獨立完成的評量目標。Andrew Ng 提出的四種代理設計模式——反思、規劃、工具使用與多代理協作——將最後一種辨識為最強大也最複雜的模式 (Ng, 2024)。在教育評量情境中，可以設想由教學代理管理教學互動並辨識評量機會、評量代理設計與施測評量任務、回饋代理生成個人化的形成性回饋，以及分析代理跨學生聚合資料以辨識學程層級的模式 (Agent4EDU, 2024)。這些代理透過結構化的通訊協定共享資訊、協調行動以避免冗餘或衝突，共同建構出比任何單一系統更為連貫、更具回應性且更全面的評量生態系統。

市場趨勢顯示，這些能力正從原型階段邁向實際部署。Gartner 預測到 2028 年，15% 的日常工作決策將由代理型 AI 自主做出；到 2026 年，40% 的企業應用將嵌入 AI 代理作為核心元件 (Gartner, 2025)。教育部門在採用曲線上通常落後於企業，但不太可能在如此規模的變革中置身事外。Yan (2025) 的 APCP 模型（Agentic-Profiling-Collaborative-Personalized），為如何在保留有意義的人類監督之下將這些能力運用於教育情境，提供了初步的藍圖。

此處有必要坦誠評估目前的證據現況。代理型 AI 在教育中的實證基礎仍處於萌芽階段。現有研究大多聚焦於生成式 AI 工具——聊天機器人、自動論文評分、適性測驗——而非能在較長時間內自主規劃、調適與行動的完全代理系統。本節關於代理型 AI 能力的論述，主要來自三類證據：技術展示與概念驗證 (Arunkumar et al., 2026; Masterman et al., 2024)、產業預測與分析師報告 (Gartner, 2025)，以及來自相鄰領域（醫療照護、軟體工程與企業自動化）的類比推理。在這些領域中，代理型 AI 的部署進展較為領先。這些來源雖確立了所述能力在技術上的可行性，但尚未構成支撐典範層級變革主張所需的嚴謹、可複製且領域專屬的實證證據。本文在此明確承認這項限制，並在全文中以審慎的保留措辭、可行性分類，以及旨在產生目前文獻尚未提供之實證的結構化試辦建議（第 5.3 節）加以處理。

語言與文化考量

另一項值得正視的實務障礙，是 AI 部署的語言面向。大多數代理型 AI 系統——包括支撐其推理能力的大型語言模型——主要以英語資料訓練而成。台灣的高等教育以華語為主要教學語言，在部分機構情境中也大量使用台灣閩南語，且許多專業學術術語在以英語為中心的訓練語料庫中可能未獲充分表徵。對於需要細緻評估論證寫作、批判思考、文化素養或以中文表達之學科專業的評量任務，以英語為主訓練的 AI 系統表現可能遠不如其英語基準所顯示的水準。

此一語言落差對 AI 增強評量的可行性與公平性均有影響。主要服務中文母語學生群體的機構，可能發現 AI 評量工具的表現不如預期可靠。以中文進行學術寫作的學生，也可能因評量系統的語言能力偏重英語而受到系統性的不利影響。因此，任何在台灣的試辦實施都必須納入對 AI 系統在中文教育情境中表現的嚴格評估。

2.3 從「**AI** 輔助」到「**AI** 代理」評量：定義邊界

前述的分類架構與能力分析，為一個具有重要實務意義的問題奠定了基礎：在教育評量的脈絡中，AI 系統何時跨越了從工具到代理的門檻？這不是純粹的理論問題。評鑑標準、機構政策、學術誠信規範與教師治理結構，都預設了一種特定的評量設計與實施模型。一旦這個模型改變——AI 從工具轉變為協作者、從器具轉變為代理——整個治理架構便需要重新審視。

表 1 綜合了三個典範類別之間的關鍵差異：傳統 AI（層級 0-1）、生成式 AI（層級 2，以作

為工具使用的大型語言模型為代表）和代理型 AI（層級 3）。

面向	傳統 AI	生成式 AI	代理型 AI
互動模式	單一查詢—回應	對話式、基於場次	自主的、目標導向的長期運作
任務範圍	狹隘、預定義的任務	廣泛但受提示詞限制	開放式、自行分解為子任務
自主性	無；完全由人類指揮	低；每次行動需人類提示	高；朝指定目標獨立運作
推論模式	基於規則或統計	單次通過式生成	迭代式、自我修正、多步驟推理
調適性	固定或限於場次內	限於對話脈絡內	跨場次、跨課程、跨時間的持久調適
輸出類型	分數、分類、標記	文本、程式碼、多媒體	行動、決策、協調性工作流程
工具使用	無	有限（插件、函式呼叫）	自主協調多種工具與系統

此一分類揭示了從生成式 AI 到代理型 AI 的轉變，並非語言模型能力的漸進式改善，而是系統與教育過程之間關係的範疇性轉移。像 ChatGPT 這樣的生成式 AI 工具在用於評量時，本質上仍是被動的：它回應人類提示、按需生成文本或評價，並在人類脫離時停止活動。相比之下，代理型 AI 系統維持自身的目標、監控自身的進展，並在沒有逐時指令的情況下持續運作——檢查學生繳交的作業、更新學習者模型、觸發介入措施。

近期一個在高等教育社群中引起廣泛討論的案例，鮮明地凸顯了此一轉變的影響。2026 年初，研究者展示了一個代理型 AI 系統（開發者暱稱為「Einstein」），能夠自主完成整門線上大學課程的所有評量，在毫無人類介入的情況下取得及格或高於平均的成績 (Inside Higher Ed, 2026)。該系統瀏覽課程管理平台、閱讀指定教材、完成測驗、撰寫論文、參與討論區並繳交期末專題。這項展示並非單純的技術噱頭，而是對當前評量實務的嚴格效度

檢驗。如果一個對教材毫無真正理解、沒有生活經驗、也沒有真實學習軌跡的 AI 代理能滿足一門大學課程的所有評量要求，那麼這些評量要求所衡量的，便可能是其所宣稱的學習成果以外的事物。「Einstein」案例挑戰的並非 AI 本身，而是未能捕捉 AI 無法複製的人類學習面向的評量設計：具身經驗、根植於個人價值觀的倫理推理、在真實社群中的協作意義建構，以及真正的智識轉化能力。

此一挑戰釐清了本文分析的核心利害。代理型 AI 不僅為評量創造新工具，它同時揭露了現有評量架構的脆弱性——這些架構從未被設計來面對具備自主、多步驟、工具使用與調適行為的代理。任何志在於代理型 AI 時代維持其適切性的品質保證架構，都必須同時處理兩個面向：善用代理型 AI 對評量的建設性潛力（第 2.2 節所述的六大能力），以及強化評量設計以抵禦這些能力所引入的效度威脅。

因此，我們提出以下操作性定義，作為後續各節分析的錨點：

> 教育評量中的代理型 **AI** 係指能自主規劃評量策略、執行多步驟評估工作流程、維持持久學習者模型，並根據累積之證據調適其方法的 AI 系統——以協作式評量夥伴而非被動工具的角色運作。此類系統以目標導向行為、工具協調能力、縱貫性記憶及多代理協作能力為特徵，且需要針對其自主決策權限（而非僅其技術輸出）加以處理的治理架構。

此定義刻意將代理型 AI 的治理意涵置於前景，因為治理層面——評鑑標準、品質保證準則、機構政策——最迫切需要調適。一個自主設計評量的系統所引發的學術權威、教師職權與機構問責問題，與一個按需批閱論文的系統根本不同。此定義同時強調協作式框架（「評量夥伴而非被動工具」），表明最具建設性的機構回應既非不加批判地採納，也非全面禁止，而是對人類教育者與 AI 代理之間的分工進行深思熟慮的重新協商。

本節所建立的分類架構與定義框架，提供了檢視具體案例的分析語彙。台灣現行的高等教育品質保證體系——其特定的評鑑準則、評量期待與機構實務——如何面對代理型 AI 的能力與挑戰？第 3 節將轉向此一問題，分析台灣評量現況中代理型 AI 既能暴露、亦有潛力改善的結構性限制。

3. 當前典範：台灣的學生學習成效衡量

任何理論化典範轉移（paradigm shift）的嘗試，都必須先嚴謹描述其所欲取代的典範。本節承擔這項任務，系統性地描繪台灣高等教育體系中學生學習成效定義、衡量與應用的制度架構。分析分為四個階段。首先，檢視高等教育評鑑中心基金會（Higher Education Evaluation and Accreditation Council of Taiwan, HEEACT）主導的評鑑架構。該架構構成體系層級學習成效評量的主要外部品質保證機制與結構性骨幹。其次，概覽教育部所部署的互補性政策工具，包括職能平台、畢業生追蹤調查及校務研究基礎設施。第三，辨識當前衡量典範中固有的六項結構性限制（structural limitations）。這些限制在 Kuhn 的術語中並非微不足道的缺陷，而是主流典範無法從自身邏輯內部解決的系統性異例。第四，追溯外部壓力的積累——人口、經濟、科技與比較性壓力——如何將這些異例加劇至危機程度。綜合而言，這些分析建立了第 4 節所理論化之典範轉移的需求條件。

3.1 HEEACT 架構：評鑑項目、核心指標與評量邏輯

台灣當代高等教育品質保證架構以 HEEACT 為核心。該組織於 2005 年成立，為受教育部委託辦理第三方評鑑的獨立非政府財團法人 (HEEACT, 2023a)。HEEACT 的成立標誌著台灣正式進入專業化外部品質保證時代，使國內實務與自 1990 年代以來重塑 OECD 各國高等教育治理的績效責任運動接軌 (Lin et al., 2021)。至 2026 年，HEEACT 已全面符合 INQAAHE 國際標準與指引，獲美國 CHEA 認可，並為亞太品質保證網絡的活躍成員 (HEEACT, 2023a; Lin et al., 2021)。

HEEACT 自 2012 年起實施雙軌制：針對所有高等教育機構的強制性外部校務評鑑，以及針對個別系所及授予學位單位的自願性系所評鑑 (HEEACT, 2024)。2017 年，教育部進行重大政策調整，將系所評鑑從強制改為自願。前提是機構須證明具備「確保教學品質之替代機制」(HEEACT, 2024, p. 2)。這項朝向機構自主與自我課責的政策轉向，代表台灣品質保證理念的重要演進——從合規導向轉為精進導向的評鑑。

校務評鑑：第三週期（***2023–2025***）

第三週期校務評鑑實施期間為 2023 年至 2025 年，預計評鑑 83 所高等教育機構。受評對象包含 67 所公私立大學、8 所宗教研修學院、6 所軍事院校及 2 所空中大學 (HEEACT, 2023a)。架構圍繞四個評鑑項目組織，各項目透過核心指標加以操作化。機構須於自我評

鑑報告中回應，並於實地訪評中展示：

- 評鑑項目一：校務治理與經營——涵蓋使命明確性、組織架構、資源規劃、決策機制、內部品質保證、校務研究及利害關係人參與（核心指標 1-1 至 1-4）。
- 評鑑項目二：教學與學術專業——處理教師表現、評鑑與獎勵制度、師資延攬品質、課程規劃與審查，以及教學品質評量（核心指標 2-1 至 2-4）。
- 評鑑項目三：學生學習與成效——與本文研究最直接相關的項目——檢視大學部教育與成效（3-1）、研究所教育與成效（3-2）、通識及跨領域教育評量機制（3-3），以及校際與跨境教育評量機制（3-4）(HEEACT, 2023a)。
- 評鑑項目四：社會責任與永續發展——涵蓋教育機會均等、社會責任實踐及財務永續（核心指標 4-1 至 4-3）。

兩個概念支柱支撐這套架構。第一，PDCA 循環融入所有評鑑項目的設計中，要求機構不僅展示機制的存在，更須提供透過迭代循環持續改善的證據 (HEEACT, 2023a)。第二，「賦權模式」引導機構整合內部品質保證與校務研究能力，允許機構在必要核心指標之外增設具特色的指標。這種設計強調機構自我精進優先於外部合規 (HEEACT, 2023a)。

評鑑流程依循五階段程序：準備、自我評鑑、書面審查、實地訪評及結果認可。每一梯次的完整流程歷時約 18 個月 (HEEACT, 2023a)。評鑑結果採三種形式之一：通過，認可效期 6 年；有條件通過，認可效期 3 年，須提交自我改善報告並接受追蹤訪評；或未通過，須重新整頓並於一年內重新申請 (HEEACT, 2023a)。這些分級結果經由評鑑委員小組與認可審議委員會的兩階段審議程序裁定，對機構聲譽具有顯著影響，並間接牽動經費分配。

系所評鑑：現行週期（***2024*** 年版）

HEEACT 系所評鑑架構最新版本為 2024 年版，與校務評鑑並行運作，層級聚焦於系所與學位學程。架構圍繞三個評鑑項目及 12 個核心指標組織：

- 評鑑項目一：學程發展、治理與改善（核心指標 1-1 至 1-4）——涵蓋教育目標、課程發展、營運效能及自我評量機制。

- 評鑑項目二：教師與教學（核心指標 2-1 至 2-4）——處理師資組成、能力建構、學術與專業發展，以及績效與機構策略的契合。
- 評鑑項目三：學生與學習（核心指標 3-1 至 3-4）——系所層級的學習成效評鑑項目，檢視學生招生管理與留校（3-1）、課程相關學習與支持系統（3-2）、其他形式的學習與支持（3-3），以及學生與畢業生學習成效及回饋（3-4）(HEEACT, 2024)。

系所評鑑結果比照校務評鑑分為三級：通過，認可效期 6 年；有條件通過，認可效期 3 年；或未通過，須重新評鑑 (HEEACT, 2024)。認可效期 3 年的學程得於 2.5 年後申請效期延長，須經額外書面審查與實地訪評。

從分析角度來看，兩個評鑑軌道如何操作化「學習成效」至為關鍵。評鑑項目三所要求的佐證資料揭示了該典範的認識論承諾：學生學習表現清單，涵蓋研究、創作、展演、實作、證照及國內外競賽獲獎；畢業後三年內的畢業生表現清單；學生學習表現的分析、檢討、回饋與改善；以及畢業生調查資料的分析、回饋與改善 (HEEACT, 2023a, pp. 32–36)。這些證據類別偏重可計量的產出與機構自我報告——第 3.3 節將檢視這種模式的限制。

3.2 教育部政策工具

在 HEEACT 評鑑架構之外，教育部部署了若干互補性政策工具，形塑台灣高等教育體系中學生學習成效的定義、衡量與激勵方式。

高等教育深耕計畫第二期（***2023–2027***）

高等教育深耕計畫（Higher Education Sprout Project）為台灣旗艦型高等教育品質提升經費機制。該計畫於 2023 年進入第二期，五年總預算約新台幣 836 億元（約 26 億美元），接續第一期（2018–2022 年）的新台幣 868.5 億元 (MOE, 2025)。計畫圍繞四大主軸：（a）強化教學品質與學習成效、（b）發展學校特色與研究卓越、（c）強化社會責任與終身學習，以及（d）促進國際競爭力。經費透過競爭性申請程序分配，機構績效依據自訂關鍵績效指標評量，其中許多直接涉及學生學習成效、畢業生就業力與雇主滿意度。

深耕計畫代表教育部透過財務誘因，將品質保證要求轉化為機構行為的主要槓桿。然而，Hou et al. (2020) 指出，該計畫的 KPI 導向評量模式可能無意間鼓勵指標最佳化，而非真正的教學轉型。機構可能傾向選擇容易衡量、滿足報告要求的指標，而非追求更深層、更難

量化的學習品質變革。

UCAN 平台

大專校院就業職能平台（University Curriculum and Career Mapping, UCAN）由教育部開發，自 2010 年起營運。平台以 Spencer and Spencer (1993) 的冰山模型為基礎，提供標準化職能評量系統。UCAN 將職業職能對應至課程結構，使機構得以評估學程與勞動市場需求的契合度。平台提供兩項主要評量工具：對應 66 個職業群集的專業職能診斷，以及涵蓋八項職場準備能力的共通職能評量。八項能力包括溝通、問題解決、創新、資訊科技應用、團隊合作、跨文化理解、生涯規劃及領導力 (MOE, 2025)。

UCAN 代表將職能導向評量引入台灣高等教育體系的重要嘗試，但其對自陳式調查的依賴及與實際課程實施的有限整合，制約了診斷效力。平台衡量的是學生對自身職能的感知，而非其展現的職能——這項區分在評估衡量典範是否充分時至為關鍵。

畢業生追蹤調查

台灣規定於畢業後 1 年、3 年及 5 年三個時間點進行系統性畢業生追蹤調查，透過教育部集中式調查系統辦理。調查收集就業狀況、薪資水準、工作滿意度、技能運用程度，以及大學教育對職涯成果的感知相關性等資料 (MOE, 2025)。調查結果同時回饋至機構自我評鑑流程與教育部政策評估。教育部期待機構分析調查結果、辨識教育供給與畢業生成果之間的落差，並實施改善。這項回饋循環記載於校務評鑑的自我評鑑報告中。

校務研究基礎設施

台灣的校務研究（Institutional Research, IR）能量處於初期至中期發展階段。台灣校務研究專業協會自 2016 年起運作，教育部也透過專項經費積極推動校務研究的建置 (MOE, 2025)。然而，各機構的校務研究成熟度差異極大。研究型大學已發展出精密的資料倉儲與分析能力，許多小型私立院校卻缺乏專責校務研究辦公室或具備量化訓練的人員。這種不均衡在當前衡量典範中產生了顯著的公平性問題——擁有最脆弱學生族群的機構，往往具有最薄弱的學習成效衡量與改善能力。

評量工具摘要

表 2 彙整台灣現行用於學生學習成效衡量的主要工具，依辦理機關、衡量類型及時間特性加以組織。

表 **2**

台灣高等教育體系學生學習成效評量工具摘要

工具	辦理機關	衡量類型	時間頻率	主要證據模式
校務評鑑（評鑑項目三）	HEEACT	機構自我評鑑的外部同儕審查	6 年週期	文件導向（自我評鑑報告＋實地訪評）
系所評鑑（評鑑項目三）	HEEACT	學程自我評鑑的外部同儕審查	6 年週期（自願）	文件導向（自我評鑑報告＋實地訪評）
UCAN 職能評量	教育部	自陳式職能診斷	每年（學生自願參與）	問卷調查（間接）
畢業生追蹤調查	教育部	自陳式就業與滿意度	畢業後 1、3、5 年	問卷調查（間接）
深耕計畫 KPIs	教育部	機構自陳式 KPI 達成	年度審查，5 年週期	量化指標＋敘述報告
課程教學評量	各高等教育機構	學生滿意度與感知學習	每學期	問卷調查（間接）
畢業專題／學位論文評量	各高等教育機構	學生作品直接評量	學程修業完成時	表現導向（直接，但未標準化）

註. 改編自 HEEACT (2023a, 2024)、MOE (2025) 及 Coates and Zlatkin-Troitschanskaia (2019)。這張表揭示一個鮮明的模式：七項主要工具中，五項依賴間接衡量（自陳報告、問卷調查、文件審查）。僅有畢業專題與學位論文評量提供學習的直接證據，且該項評量在機構層級辦理，缺乏標準化或跨校可比較性。間接衡量的主導地位構成當前典範的基礎特徵。

3.3 當前典範的結構性限制

描繪制度架構之後，接下來辨識其中嵌含的結構性限制。這些限制並非可藉漸進式改革補救的偶發缺陷。它們在 Kuhn (1962/2012) 的術語中是異例（anomalies）——典範承諾衡量的內容與實際捕捉的內容之間存在系統性落差。以下辨識出六項異例。

時間性限制（**temporal limitation**）。當前典範以週期性快照邏輯運作。校務評鑑遵循 6 年週期，系所評鑑同樣在多年窗口中評量表現 (HEEACT, 2023a, 2024)。自我評鑑報告回顧過去，將 3.5 至 4 年的歷史資料濃縮為一份在實地訪評前數月即提交的文件。畢業生追蹤調查則捕捉畢業後 1 至 5 年的就業成果。在快速技術變革、勞動市場波動與教學創新日益形塑高等教育的環境中，這種時間架構使評量系統在結構上成為回顧導向的 (Coates & Zlatkin-Troitschanskaia, 2019)。一個學程可能基於生成式 AI 出現前畢業的世代成果而獲得通過，卻無任何機制可評量現行學生是否正在為後 AI 時代的勞動市場做準備。

粒度限制（**granularity limitation**）。評量架構在機構與學程層級運作，產出關於集體實體的判斷——「該機構通過評鑑」、「該學程符合評鑑項目」。這些總體性判斷遮蔽了個別學習軌跡。一個學程可能通過評鑑，其中的個別學生卻無形地掙扎著。一所大學可能在自我評鑑報告中展示令人滿意的學習成效，但特定次群體（第一代大學生、身心障礙學生、社經弱勢背景學生）卻經歷顯著不同的教育現實。當前典範並無任何機制能以系統性、即時的方式，將學習成效細分至學程層級以下 (Coates & Zlatkin-Troitschanskaia, 2019)。這種粒度落差不僅是衡量上的不便，而是分析單位（學程與機構）與教育關懷單位（個別學習者）之間的根本性錯配。

模態限制（**modality limitation**）。當前典範中占主導地位的證據模態為文件式。自我評鑑報告構成評鑑判斷的主要依據，輔以為期 1 至 2 天的實地訪評 (HEEACT, 2023a)。這種文件導向途徑無法捕捉即時學習歷程、逐時的教學互動，或學生職能在整個學期或學位修業期間的動態演進。評量根本上依賴機構所報告的內容，而非學生所經歷的內容。自我評鑑報告就其本質而言是一份策展性敘事——機構挑選支持有利評價的證據，按照規定架構加以組織，並以最適於評鑑委員評量的格式呈現。這並非批評機構誠信，而是指出文件模態本身在證據基礎中引入了系統性選擇偏誤 (Tam, 2001)。

代理限制（**agency limitation**）。在當前典範中，被評量的實體同時也是報告者。機構自行

準備自我評鑑報告、選擇證據、定義特色指標，並呈現教育品質的敘事。實地訪評雖提供部分外部查核，HEEACT 評鑑委員也在評估自陳性主張時運用專業判斷，但根本的資訊不對稱仍然存在。機構對自身實際表現所掌握的資訊，遠多於任何文件審查或 1 至 2 天訪評所能捕捉的。這種結構性特徵造就了組織理論學者所稱的「印象管理」誘因 (Goffman, 1959)——未必不誠實，但系統性地傾向有利的自我呈現。代理限制在學習成效衡量上尤為明顯，因為報告性成果（獎項、證照及就業率清單）與實際學習（批判性思考發展、深度理解、職能成長）之間的落差在此最為巨大。

職能捕捉限制（**competency capture limitation**）。 HEEACT 評鑑項目中規定的證據類別，揭示了對可計量產出優先於複雜職能的系統性偏好。校務評鑑項目三所要求的佐證資料強調學生學習表現清單——研究產出、創作與展演、實作、證照，以及國內外競賽獲獎——連同畢業生調查資料 (HEEACT, 2023a, pp. 32–36)。這些產出是學習的合理證據，但僅代表 Spencer and Spencer (1993) 冰山模型中可見的冰山一角：容易觀察且可衡量的知識與技能。在水面下——自我概念、特質與動機等更深層、更具預測力的職能層次——當前典範並無系統性的衡量方法。批判性思考、創意問題解決、跨文化敏感度、倫理推理及協作智慧等複雜職能，恰恰最受雇主重視，也最抗拒當前架構的產出計量邏輯 (Association of American Colleges and Universities [AAC&U], 2018)。

間接衡量主導（**indirect measurement dominance**）。如表 2 所記錄，當前典範壓倒性地依賴學習的間接證據——問卷調查、自陳報告、文件審查——而非依據既定學習成效對學生表現進行直接評量。UCAN 衡量的是感知職能；畢業生追蹤調查衡量的是自陳就業與滿意度；課程教學評量衡量的是學生對教學的感知；評鑑本身評量的是機構用以評估學習的機制，而非學習本身。唯一具一致性的直接評量工具——畢業專題、碩博士論文——在機構層級運作，缺乏標準化、跨校可比較性，或對學習歷程隨時間進展的系統性分析。這種模式反映了國際文獻中辨識的更廣泛挑戰：全球高等教育體系在建構品質保證流程方面，比在發展穩健的學生學習衡量工具方面更為成功 (Coates & Zlatkin-Troitschanskaia, 2019; Shavelson, 2010)。

3.4 異例的積累：當前典範為何承受壓力

上述六項結構性限制自台灣現代品質保證架構建立之初即已存在。改變的是外部環境——它將這些限制從可容忍的不完美加劇為迫切的異例。若干匯聚的壓力正將當前典範推向崩潰的臨界點。

人口危機

台灣的出生率為全球最低之列，已為高等教育部門帶來存亡等級的招生危機。高等教育在學總人數約為 90 萬人，教育部預估大一新生註冊人數將於 2030 年代初期降至約 173,000 人——許多機構無法承受這個數字 (MOE, 2025)。僅 2024 年即有 7 所大學停辦，預測顯示至 2028 年可能有多達 40 所私立大學退場 (Sharma, 2024; Taiwan News, 2024)。人口現實轉變了品質保證的核心問題：在緊縮的體系中，問題不再僅僅是「機構是否達到最低門檻？」而是「評量系統能否辨識哪些機構正在提供足夠的學習價值以證明其持續營運的正當性？」更關鍵的是，系統能否足夠迅速地做到，以保護瀕臨退場機構中的學生？6 年評鑑週期在結構上無法提供這種預警功能。

就業失配

台灣大學就學率超過 95%，高等教育體系實質上已普及化。然而畢業生失業率約為 4.5%，更重要的是，愈來愈多證據顯示畢業生職能與雇主需求之間存在質性失配 (MOE, 2025)。一項在 PACIS 發表的近期研究發現，台灣 63% 的 AI 相關職位要求應用層級技能，而僅 37% 的高等教育課程涵蓋這些職能。這項結構性落差是現行評量機制未能及時偵測或修正的。就業失配對當前典範的正當性破壞尤為嚴重，因為畢業生就業成果正是該典範確實衡量的少數指標之一。然而即便在此處，衡量時間過於延遲（畢業後 1 至 5 年）且粒度過於粗略（總體就業率），無法驅動有意義的課程調適。

教師對成果導向教育的抵制

國際上朝向成果導向教育的轉向——HEEACT 的評鑑架構透過在所有評鑑項目中強調學習成效而正式認可——在教師層級面臨顯著的實施阻力。針對台灣高等教育的研究記錄了教師對成果導向教育的普遍懷疑。懷疑根源在於對學術自由的擔憂、職能架構的化約主義、成效文件撰寫的行政負擔，以及對於教育目的能否或應否以可衡量成效加以捕捉的根

本性哲學歧見 (Lin et al., 2021)。這種抵制在形式架構（要求成效證據）與基層現實（教學與評量實務可能仍以投入與過程為焦點）之間製造了落差。當前典範除了自我評鑑報告中經過修飾的證據之外，並無機制偵測或處理這種實施落差。

數位落差

台灣高等教育體系涵蓋巨大的機構多樣性——從擁有精密數位基礎設施的研究密集型國立大學，到科技能量有限的小型私立院校。當前評量典範採文件導向，對數位落差保持不可知論。它透過相同的自我評鑑報告與實地訪評流程評量所有機構，不論其科技成熟度。然而，隨著數位學習環境日益成為教育實施的核心（因 COVID-19 疫情而加速），評量架構無法評估數位教學法、線上學習品質或教育科技效能，代表一個日益擴大的盲點。

國際比較對象進展更快

台灣的品質保證架構並非孤立存在，它運作於日益競爭的國際格局中，而同儕體系正在快速演進。歐洲高等教育品質保證標準與指引於 2015 年進行重大修訂，目前正進行進一步更新，日益整合數位職能與學習分析的考量 (ENQA, 2015)。澳洲 TEQSA 已制定明確的數位教學品質保證要求，並正探索 AI 增強的品質評量流程。新加坡於 2023 年宣布的教育科技總體規劃 2030，將科技增強的學習評量定位為國家策略優先事項 (MOE Singapore, 2023)。對照這些比較對象，台灣以文件導向、6 年週期、大量依賴問卷調查的評量典範面臨落後風險。原因並非缺乏嚴謹性，而是它所提供的嚴謹性屬於一種對新興高等教育環境而言正變得結構性不足的類型。

AI 職能落差

或許最具後果性的異例同時也是最新的：當前評量架構並無機制將 AI 素養作為學生學習成效加以評量。在生成式 AI 正快速轉變幾乎所有領域專業實務的時代，評鑑項目、UCAN 評量模組及畢業生追蹤工具中 AI 職能的缺席，意味著該典範在結構上無法評量台灣畢業生是否已為畢業後將面對的專業現實做好準備。這不僅僅是可藉新增一項指標即可填補的內容缺口。它代表一項典範層級的挑戰，因為 AI 職能並非可透過傳統產出計量衡量的靜態知識體系，而是一種動態、持續演進的能力，需要全新的評量模態方能衡量。

綜合而言，這些積累的異例創造了 Kuhn (1962/2012) 所辨識的、先於典範轉移而出現的「危機」條件。當前典範——建立於週期性循環、文件導向證據、機構自我報告、總體性判

斷及間接衡量之上——在高等教育擴張、變革漸進、品質保證的首要問題為機構是否達到最低門檻的時代是足夠的。但那個時代正在結束。台灣現今面對的是一個伴隨加速變革而緊縮的體系、一個要求現行工具無法衡量之職能的勞動市場、一場需要 6 年週期無法提供之預警能力的人口危機，以及一場使整個評量模態可能過時的 AI 革命。問題不在於當前典範是否會改變，而在於如何改變——以及台灣能否主動而非被動地引導這場變革。

第 4 節提出一個答案：將代理型 AI 系統整合至學習成效衡量架構中。這不僅僅是技術升級，而是在學生學習的定義、捕捉、分析與行動上的典範性轉型。

4. 典範轉移：從靜態測量到動態學習證據

前述章節已建立兩項基礎論述：代理式人工智慧（agentic AI）所具備的能力組合——規劃、適應、工具使用、多步驟推理、記憶與多代理人協作——為教育評量創造了質性上全新的可能性（第二節）；而台灣現行品質保證典範展現出六項結構性限制，無法透過漸進式改良加以解決（第三節）。本節綜合上述論述，以回應本文的核心研究問題：代理式人工智慧透過何種機制，能將學習成果測量從定期性、標準化、總結性的取徑，轉型為持續性、個人化、形成性的典範？援引 Kuhn (1962/2012) 的科學革命理論，本節提出 ADAPT 框架——Assessment-Design for Agentic Paradigm Transformation（代理式典範轉型之評量設計）——作為理解此一轉型的原創性概念模型，闡述典範轉移運作的七大面向（dimensions），並透過一個以台灣科技大學為場景的情境範例來具體說明新典範。關鍵的是，本節亦直面技術轉型的局限，檢視人類判斷不可化約的角色，以及重新定義何謂學習證據所帶來的本體論意涵。

4.1 Kuhn 的框架應用於評量理論

Thomas Kuhn (1962/2012) 對科學革命的論述，為理解當前教育評量的處境提供了一個有力的分析視角。Kuhn 主張，科學領域運作於典範（paradigm）之中——一套共享的假設、方法與範例框架，界定了何為正當的問題與可接受的解答。在典範內的進展構成「常態科學」（normal science），但當典範無法解決的異例（anomalies）持續累積，便會出現危機（crisis），最終催化出一場革命性轉移，進入與舊典範不可共量（incommensurable）的新典範。

為品質保證採用 ***Kuhn*** 分析框架的正當性

Kuhn 的框架最初是為自然科學所發展的，他本人也曾表達對將其延伸至社會領域的懷疑。這種懷疑必須被嚴肅對待。然而，大量學術文獻已證明 Kuhn 的核心概念——典範、常態科學、異例、危機與革命——在經過適當限定後，其分析效用遠超自然科學範疇。Masterman (1970) 在其極具影響力的分析中，辨識出 Kuhn 使用「典範」一詞的二十一種不同涵義，並主張其社會學意涵——一套共享的範例、實踐與標準，界定了一個專業社群的工作方式——適用於任何有組織的實踐領域，而非僅限於科學學科。Ritzer (1975) 將 Kuhn 的分析應用於社會學，證明典範概念能夠闡明專業社群如何採納、捍衛並最終放棄共享的實踐框架。Eckstein (1992) 則將此框架延伸至政治學，顯示當行政與管理體制展現出 Kuhn 所辨識的關鍵特徵——共享的範例、由被接受方法所構成的「學科矩陣」（disciplinary matrix），以及界定正當問題與解答的專業社群——時，便能作為典範運作。

本文將 Kuhn 的框架作為分析性啟發工具（analytical heuristic），而非嚴格的知識論主張。我們的論點並非品質保證是一門正在經歷 Kuhn 精確意義上之科學革命的自然科學，而是 Kuhn 的語彙對理解當前教育評量的處境具有分析上的生產力。台灣的品質保證系統展現了使 Kuhn 類比具有啟發性的結構特徵：它透過共享範例運作（自我評鑑報告範本、實地訪視協定、核心指標架構）；具有一套學科矩陣（結構化所有評鑑活動的 PDCA 循環）；擁有一個具備明確規範的專業社群（HEEACT 的評鑑委員團隊，經過特定方法論訓練並共享證據標準）；以及一組公認的「謎題」（如何測量學習成果、如何確保機構問責），由從業者在典範邏輯內加以解決。這些特徵並非隱喻性的，而是可觀察到的制度結構，其運作方式類同於 Kuhn 的典範組成要素。

「典範革命」的框架設定與本文建議採用情境 B（漸進式框架演化）之間的表面張力，需要直接面對。這種張力比表面上看來更不矛盾。Kuhn 本人即已承認——特別是在《科學革命的結構》1969 年版後記中——典範轉移並非總是突然的斷裂；在應用領域中，隨著從業者累積舊框架不足的證據並逐步採用新方法與標準，典範轉移可以漸次展開。情境 B 代表的正是 Kuhn 所描述的「過渡時期」（transition period）——一個新舊典範要素共存的階段，從業者甚至在完整的格式塔轉換（gestalt shift）發生之前，便已開始重新概念化其工作。第 5.3 節所提出的三階段實施路徑，旨在刻意管理這一過渡，既不強迫過早的革命，也不

排除隨著證據基礎成熟而實現典範層級變革的可能。

更廣泛地說，學者們已將 Kuhn 的框架富有成效地應用於教育領域 (Shepard, 2000)，以及更晚近地，應用於人工智慧如何重塑教育典範的具體問題 (Zhong & Zhao, 2025)。

評量中的常態科學。當前高等教育品質保證典範——定期性、標準化、總結性、文件導向、機構自報——構成了 Kuhn 所認定的常態科學。其方法已然確立：五至七年的評鑑週期、標準化的自我評鑑報告、在學程或機構層級彙總的統計指標，以及由人類小組進行的同儕審查 (HEEACT, 2023)。在這個典範之內，品質保證專業人員精煉工具、調整指標權重、開發評量規準——這正是常態科學的「解謎」（puzzle-solving）活動 (Kuhn, 1962/2012, p. 35)。這個典範已產出真正的成就：台灣的 HEEACT 已建構出亞洲最受推崇的品質保證系統之一，定期評鑑已可證實地提升了機構對教學品質與學生成果的關注 (Hou et al., 2012)。

異例。然而，第三節所辨識的六項結構性限制——時間僵固性（temporal rigidity）、粒度塌縮（granularity collapse）、模態限制（modality restriction）、能動性不對稱（agency asymmetry）、能力捕獲失靈（competency capture failure），以及間接測量主導（indirect measurement dominance）——正如 Kuhn 式異例般運作。它們代表了現行典範認為重要但無法在其自身邏輯內解決的問題。例如，時間性限制並非可以透過將評鑑週期從六年縮短為三年來修復的缺陷；即使是年度評量，本質上仍是定期性而非持續性的。粒度限制也不是透過蒐集更多彙總資料就能解決的；它需要一個完全不同的分析單位。這些不是等待在現有典範中以更巧妙方案來解決的謎題，而是典範本身所固有的結構性矛盾。

危機。三項外部壓力將這些異例加劇為 Kuhn (1962/2012) 所稱的危機。首先，台灣的人口衰退——大學一年級新生從 2015 年的約 27 萬人降至 2028 年預估的 17.3 萬人 (Ministry of Education, 2024)——使得每位學生的學習軌跡都攸關機構存亡，彙總性的世代指標因此變得危險地不足。其次，儘管經過數十年的評鑑，雇主對畢業生能力的不滿持續存在，這暗示現行典範測量的是機構合規性而非真實的學習成果 (MOE, 2025)。第三，AI 能力作為畢業生必備素養的快速崛起，創造了現有工具從未被設計來捕獲的評量需求——針對創造力、人機協作、演算法脈絡中的倫理推理 (UNESCO, 2025)。如同 Kuhn 的框架所預測，這些異例的累積重量產生一種感覺：典範本身存在根本性問題，而非僅是其實施方式有問

題。

革命。代理式人工智慧並非僅對現有典範提供漸進式改進——更快速的資料處理、更有效率的報告產出或自動化的指標計算。它所實現的，是 Kuhn (1962/2012) 所描述的格式塔轉換：一種根本不同的評量觀看方式。現行典範將學習視為一個在定期間隔測量的端點，代理式典範則將學習視為一條持續被觀察的軌跡。現行典範將證據視為機構產出的文件，代理式典範則將證據視為從學習過程本身湧現的多模態痕跡（multimodal traces）。現行典範將評量者定位為外部稽核員，代理式典範則構想評量為學生、教師、AI 代理人與品質保證機構之間的協作活動。這些不是量化差異——更多資料、更快處理——而是評量本體論（ontology of assessment）上的質性差異：學習是什麼、什麼證據算數、誰來評量。如 Zhong and Zhao (2025) 所論，AI 時代要求對教學、學習與評量典範進行根本性的反思。

不可共量性。 Kuhn (1962/2012) 最具爭議的主張——連續的典範之間不可共量，意即它們無法被完全翻譯為彼此的術語——在此具有特殊的適用力。畢業率與學習軌跡並非同一事物的不同測量方式；它們反映的是對「學生成功」意味什麼的根本不同概念。評鑑小組的判斷與 AI 代理人的持續性能力繪圖並非回答同一問題的不同方法；它們回答的是完全不同的問題。這種不可共量性並不意味新典範使舊典範變得毫無價值——定期的人類審查仍保有其本質價值，如第 4.5 節所論——但它確實意味兩個典範無法被平滑地混合。此一過渡需要刻意的架構性抉擇，而 ADAPT 框架正是為支持此種抉擇而設計。

4.2 ADAPT 框架：一個轉型概念模型

為提供一個結構化的分析視角，以理解代理式人工智慧如何轉型學習成果測量，本文提出 ADAPT 框架：Assessment-Design for Agentic Paradigm Transformation（代理式典範轉型之評量設計）。此框架並非規範性的實施指南，而是一個概念模型，使機會與風險的系統性分析成為可能。

推導邏輯

ADAPT 框架的五個面向源自前述章節所發展之三項分析輸入的交叉點：技術能力分析（第二節），辨識代理式人工智慧能做什麼；診斷性限制對應（第三節），辨識現行典範無法做什麼；以及在教育中負責任 AI 部署文獻中所辨識的治理要求 (UNESCO, 2023; Beauchamp & Childress, 2019)。本框架的面向並非任意選取或對本文研究問題的重新標記；

它們代表了將技術能力連結至機構轉型、同時維持倫理治理所需的最小分析類別集合。理解代理式人工智慧能做什麼（Agency Architecture，能動性架構）是辨識現行典範何處失靈（Diagnostic Mapping，診斷對應）的先決條件；這些診斷性缺口界定了評量實踐中什麼必須改變（Assessment Reconception，評量再概念化）；經過再概念化的評量需要制度與法規的調適（Policy Pathways，政策路徑）；而上述所有面向都必須受到倫理防護機制的約束（Trust & Ethics Safeguards，信任與倫理防護）。

關係結構

不同於一份檢核表，ADAPT 框架主張這五個面向具有順序依賴性（sequentially dependent）且相互約制（mutually constraining）——若僅處理單一面向而未觸及其他面向，便有產出技術上可行但制度上不可行或倫理上不可接受之方案的風險。其關係結構運作如下：Agency Architecture（A，能動性架構）揭示使新評量模態成為可能的技術能力；此一理解進入 Diagnostic Mapping（D，診斷對應），將這些能力與現行典範中的具體結構性限制進行匹配。診斷缺口一旦被辨識，便界定了 Assessment Reconception（A，評量再概念化）所必須達成的目標——即第 4.3 節闡述的典範轉移七面向。經過再概念化的評量，接著需要 Policy Pathways（P，政策路徑）將概念可能性轉譯為法規與制度現實。而政策的實施則要求 Trust & Ethics Safeguards（T，信任與倫理防護）來約束什麼是可允許的——並非所有技術上可行或診斷上有依據的事物在倫理上都是可接受的。關鍵的是，此框架具有遞迴性（recursive）：信任考量約束了哪些能動性架構是可允許的（T 回饋至 A），而診斷對應揭示了政策路徑在何處必須最為審慎地設計（D 約束 P）。ADAPT 框架因此同時作為線性分析進程與相互約制的回饋迴圈（feedback loop）運作，反映了轉型一個既有品質保證典範的複雜性。

ADAPT 框架包含五個相互關聯的組成要素：

A — Agency Architecture（能動性架構）處理的是在評量脈絡中 AI 何以具有「代理性」，以及第二節所辨識的六項能力——規劃、適應、工具使用、多步驟推理、記憶與多代理人協作——如何結合以創造出具備真正自主性的評量系統。此組成要素對應研究問題一（RQ1），並建立典範轉移的技術基礎。

D — Diagnostic Mapping（診斷對應）將現行典範的結構性限制（第三節）連結至 HEEACT

評鑑框架中的具體標準，精確辨識現行典範在何處失靈，以及代理式人工智慧的能力在何處提供轉型性替代方案。此組成要素對應研究問題二（RQ2），並將框架紮根於台灣的具體制度脈絡。

A — Assessment Reconception（評量再概念化）闡述典範轉移運作的七個面向（第 4.3 節），從靜態測量模型邁向動態學習證據模型。這是本框架的核心分析貢獻，對應研究問題三（RQ3），並提供了描述什麼改變以及為何改變的詞彙。

P — Policy Pathways（政策路徑）辨識台灣品質保證框架可能演化的三種情境——保守整合、適度轉型與典範革命——各有其對法規設計、機構能量與學生公平性的不同意涵。此組成要素對應研究問題四（RQ4），並在第五節詳加闡述。

T — Trust and Ethics Safeguards（信任與倫理防護）繪製典範轉移各面向之伴隨風險，並提出治理機制——包括演算法透明性、公平性稽核、同意架構與人類覆議協定——這些機制必須伴隨技術轉型。此組成要素對應研究問題五（RQ5），並在第六節詳加闡述。

本框架的分析效用恰恰在於此種關係性架構：它不僅組織了本文的論述，更產生分析性預測——例如，若處理評量再概念化卻未有相應的政策路徑調適，將產出技術上精緻但制度上不可行的提案；或者，在缺乏信任防護的情況下部署代理式能力，將產生類似南韓 AI 教科書回撤事件般的利害關係人反彈（詳見第 5.2 節討論）。

4.3 典範轉移的七個面向

ADAPT 框架中的 Assessment Reconception（評量再概念化）組成要素，辨識出典範從靜態測量轉向動態學習證據所運作的七個面向。表 3 摘述了這些面向；隨後的分析逐一詳加檢視。為維持認識論上的精確性，討論中區分三類能力：已證實能力（demonstrated capabilities），存在於當前教育技術實施中（例如，自動論文評分、適性測驗、學習分析儀表板）；新興能力（emerging capabilities），已在原型或有限部署中被證明（例如，多模態評量、跨課程能力追蹤）；以及預期能力（projected capabilities），技術上可行但尚未在大規模教育情境中被證明（例如，自主多代理人評量生態系統、即時彙總跨學程學生層級資料的機構品質儀表板）。

表 **3**

評量典範轉移的七個面向

面向	現行典範	代理式 AI 典範	變革機制
時間性（Temporality）	定期性（5-7 年週期）	持續性	持久記憶 + 全時監測
粒度（Granularity）	彙總性（學程/世代）	個體軌跡	動態學生建模 + 個人化評量
能動性（Agency）	機構自報	AI 觀察 + 學生共創	自主資料蒐集 + 多代理人證據綜合
回饋延遲（Feedback latency）	回溯性（數月/數年）	即時性	迭代推理 + 即時回饋迴圈
證據類型（Evidence type）	文件、統計、問卷	多模態學習痕跡	工具使用 + LMS 整合 + 多模態分析
評量目的（Assessment purpose）	問責/合規	改善/個人化	目標導向規劃 + 適性介入
評量者（Assessor）	人類審查者	人機協作評量	多代理人協作 + 人在迴圈中監督

4.3.1 時間性（***Temporality***）：從定期到持續

在現行典範下，學習成果在固定間隔測量——通常對齊於五至七年的評鑑週期——產出的是在被分析之前可能已經過時的時間快照（第 3.1 節）。相較之下，代理式 AI 系統可以潛在地利用持久記憶（persistent memory）來維持持續更新的學習過程模型。不同於回應個別提示的無狀態 AI 工具，代理式系統跨互動保留脈絡，建構在學期與年度間演化的累積性學生學習表徵 (Arunkumar et al., 2026; Masterman et al., 2024)。

在台灣的大學脈絡中，這意味著嵌入學程學習管理系統的 AI 代理人會持續監測學生在各課程間的表現，追蹤的不僅是成績，還包括參與模式、概念發展以及能力習得。代理人的

持久記憶使其能偵測到縱貫性趨勢——例如，一個世代的分析寫作品質逐漸下滑——而定期評量要在數年後才能捕獲這種已經開始的衰退。OpenAI (2025) 的學習成果測量工具展示了此類持續監測的早期原型，儘管目前的實施距離這個面向所構想的無縫整合仍有相當距離。

伴隨此一轉移的挑戰是監控蔓延（surveillance creep）。持續監測，即使意圖良善，也會創造無所不在的觀察環境，可能抑制冒險、創造力，以及對深度學習至關重要的建設性失敗（productive failure）(Bearman & Ajjawi, 2023)。因此，典範轉移的時間性面向需要審慎校準：在能力上持續，但在應用上審慎，對於監測什麼、何時監測、以什麼目的監測設定清晰的界線。

4.3.2 粒度（***Granularity***）：從彙總到個體

現行品質保證以學程和世代為運作單位：平均畢業率、彙總就業統計、平均滿意度分數。這些彙總指標雖有助於機構間標竿比較，卻遮蔽了群體內的變異——而這些變異往往比平均值更具資訊量 (Pellegrino et al., 2001)。代理式 AI 典範將分析單位轉移至個體學習軌跡（individual learning trajectory），運用動態學生建模來建構個人化的能力檔案。

此一轉移由適應性行為的技術先決條件所創造：AI 代理人可根據個別學生的反應、學習模式與已展現的能力調整其評量策略，如同一位專家導師 (Black & Wiliam, 1998)。對台灣的大學而言，這意味著從「本學程畢業生中有多少百分比在一年內就業？」這個問題，邁向更為豐富的問題：「這位學生的問題解決能力在四年間如何發展？哪些學習經驗最為關鍵？」

此處的風險是公平性（equity）。由 AI 驅動的個人化評量，若模型從訓練資料中編碼了偏見、若數位素養較高的學生產生更豐富的學習痕跡、或者若資源匱乏的機構無法實施精密的 AI 系統，則可能加劇既有的不平等。Banihashem et al. (2025) 提醒，為形成性評量所最佳化的學習分析可能無意間創造出「回饋豐富」與「回饋貧乏」的環境，鏡射並放大既有的社會經濟差距。

4.3.3 能動性（***Agency***）：從機構自報到 ***AI*** 觀察與學生共創

或許最深刻的面向轉移涉及的是誰產出評量證據。在現行典範下，機構策展自我評鑑報告，選擇呈現哪些資料以及如何框架——這種安排製造了固有的利益衝突（第 3.4 節）。代

理式人工智慧引入兩個新的證據來源：自主 AI 觀察與學生共創。

具備工具使用能力的 AI 代理人可獨立存取學習管理系統、作業資料庫與機構資料庫，產出未經機構自報篩選的證據 (Ng et al., 2021)。同時，學生成為其評量證據的積極共創者，與 AI 代理人合作建構反映其對自身學習成長理解的能力歷程檔案（competency portfolios）。此種雙重轉移——從機構壟斷到三角驗證的證據生態系統——回應了第三節所辨識的能動性不對稱問題，同時契合將學習者定位為主動意義建構者而非被動測量對象的建構主義原則 (Biggs, 1999)。

伴隨的挑戰是課責性（accountability）。當證據由 AI 代理人而非人為撰寫的報告所產出時，便出現了關於錯誤責任歸屬、AI 產出證據的可解釋性，以及演算法產出之評量的法律與法規地位等問題。當 AI 代理人的能力評量與教師的專業判斷相矛盾時，誰該負責？

4.3.4 回饋延遲（***Feedback Latency***）：從回溯到即時

現行典範的回饋迴圈以數月至數年的時間尺度運作：評鑑報告提交、審查、討論，並在被評量的學習發生很久之後才傳達回機構（第 3.1 節）。代理式人工智慧的迭代推理能力——將複雜評量分解為可管理的步驟、依序執行，並根據中間結果精煉結論——使評量回饋得以趨近即時 (Masterman et al., 2024)。

此處的機制不僅是更快速的運算，而是根本不同的評量架構。代理式系統並非在週期結束時批次處理評量，而是持續產出可以被立即採取行動的形成性回饋。Black and Wiliam (1998) 的經典後設分析（meta-analysis）證明，附帶即時回饋的形成性評量是改善學習最有力的介入措施之一；代理式人工智慧提供了在規模上、持續地、並針對個別學習者個人化地提供形成性評量的基礎設施。

在台灣的脈絡中，這可能意味著學程主管收到一則警示——不是在下一次評鑑報告中，而是在下週二——告知某門課程中的學生正在某個門檻概念（threshold concept）上遭遇困難，同時附有 AI 代理人對可能原因的分析以及建議的介入措施。從回溯性報告到即時情報的轉變，將品質保證從一個向後看的問責活動轉型為一個向前看的改善引擎。

風險是資訊超載（information overload）。即時回饋系統可能產出超過人類能有意義地處理的資料量，導致警示疲勞（alert fatigue），並弔詭地使機構回應能力降低 (Swiecki et al., 2022)。有效的實施需要 AI 代理人能對何時升級做出判斷——這是一種仰賴代理式系統之

規劃與適應功能的能力。

4.3.5 證據類型（***Evidence Type***）：從文件到多模態學習痕跡

現行品質保證高度依賴文字文件——自我評鑑報告、課程大綱、會議記錄——並輔以結構化統計資料（第 3.3 節）。代理式典範將證據基礎擴展為涵蓋多模態學習痕跡（multimodal learning traces）：在學習過程中生成的數位製品（digital artifacts），包括程式碼儲存庫、設計歷程檔案、實驗室筆記、影片報告、協作文件、同儕審查交流以及反思日誌。

此一擴展在技術上得以實現，是透過代理式 AI 的工具使用能力——與多元數位系統介接、擷取異質資料並將其綜合為連貫的證據敘事的能力 (Ng et al., 2021)。AI 代理人並非依賴學生透過標準化考試來展現其學習，而是在學習跨多種模態與環境展開的過程中進行觀察。Shute and Ventura (2013) 將此取徑稱為「隱匿式評量」（stealth assessment）——從學生與學習環境的自然互動中提取其能力證據，而不以正式測驗中斷那些互動。隱匿式評量的價值與第 6.1.1 節所闡述的知情同意要求之間存在張力：在學生不知情的情況下進行評量，即使出於教學動機，也有違反本文所支持之自主原則的風險。一個「透明隱匿式評量」（transparent stealth assessment）的設計原則可以調和此一張力——事先告知學生其自然學習互動將產出評量證據，且學生理解所涉及的一般機制，但證據提取的具體時刻與模態不會即時示意，從而保留使隱匿式評量具有診斷價值的生態效度（ecological validity）。

對台灣的大學而言，多模態證據蒐集可以回應長期以來對標準化評量未能捕獲專業實踐中最重要能力的批評。例如，一位護理系學生的能力，透過她在模擬臨床遭遇（simulated clinical encounters）中的表現、反思實踐日誌，以及在團隊照護情境中的協作問題解決，遠比透過一場紙筆測驗的分數更能被有效地證明。代理式人工智慧提供了將這些多元證據流綜合為連貫能力檔案的整合基礎設施。

挑戰在於效度（validity）。多模態學習痕跡是嘈雜的、脈絡化的且難以標準化的。要確立 AI 代理人對多元證據流的綜合構成有效的評量——即它確實測量了它所宣稱測量的東西——需要超越古典測驗理論（classical test theory）的新評量效度框架 (Mislevy et al., 2003)。以證據為中心的設計（Evidence-Centered Design, ECD），透過其對證據宣稱、任務情境與評分規則的明確建模，提供了一個有前景的基礎，但必須加以大幅擴展，以容納代理式 AI 動態、多模態的證據生成。

建構效度的挑戰

證據類型的擴展引發了一個必須直接面對的根本心理計量關切。AI 生成之能力評量的核心挑戰是建構效度（construct validity）：AI 系統在多大程度上測量的是目標建構（例如批判性思考、倫理推理、協作問題解決），而非與該建構相關的替代指標（例如論文長度、關鍵詞使用、回應時間、LMS 登入頻率）。此一關切並非 AI 所獨有——人類評量者有時也會混淆替代指標與建構——但在從訓練資料中學習統計關聯而缺乏使人類評量者能區分真正能力與表面表現之脈絡理解的 AI 系統中，這一問題被放大了。

與 AI 增強評量最為相容的心理計量框架是 Mislevy et al. (2003) 所闡述的以證據為中心的設計（Evidence-Centered Design, ECD）。ECD 明確建模三個相互連結的組成要素——能力模型（competency model，評量什麼）、證據模型（evidence model，哪些可觀察行為構成能力的證據）與任務模型（task model，哪些任務引發那些可觀察行為）——為設計能闡明為何特定觀察能算作特定能力之證據的 AI 評量系統，提供了原則性的基礎。缺乏這樣的框架，AI 系統便有產出統計上可靠（跨施測一致）但不具效度（並未實際測量其所宣稱測量之事物）之評量的風險。

AI 生成評量的效度證據仍處於起步階段。自動論文評分系統已累積了二十餘年的效度證據，證明在特定評量脈絡中（例如，具有明確評量規準的標準化寫作評量）與人類評分者具有可接受的一致性。然而，將 AI 評量延伸至複雜的、非結構化的能力——正是代理式典範所致力於評量的能力——尚缺乏可比擬的驗證。因此，本文提議所有 AI 增強評量的試點實施（第一階段，第 5.3 節）應將系統性效度研究納為必要組成要素。這些研究不應僅檢視 AI 與人類評量之間的統計一致性，更應檢視更深層的問題：AI 生成的能力檔案是否對應於學生學習中有意義的差異——這個問題只能透過縱貫追蹤與對照多重獨立測量的聚斂驗證（convergent validation）來回答。

4.3.6 評量目的（*Assessment Purpose*）：從問責到改善

現行典範壓倒性地是總結性的：其主要目的是為問責目的對機構品質做出判斷——評鑑決定、排名、經費配置 (HEEACT, 2023)。雖然總結性評量服務於正當的社會功能，但其主導地位排擠了直接改善學習的形成性評量用途 (Shepard, 2000)。代理式典範將重心從問責轉向改善、從判斷轉向發展、從總結性裁定轉向形成性情報。

此一轉移之所以可能，是透過代理式 AI 系統的目標導向規劃能力。代理式評量系統並非被動地記錄結果以供日後判斷，而是可以主動規劃介入措施：辨識高風險學生、推薦教學調整，並產出針對特定學習落差量身訂做的資源 (Banihashem et al., 2025)。AI 代理人不僅測量；它根據測量結果採取行動，並以改善學習成果為明確目標。

這所產生的張力存在於促進學習的評量（assessment for learning）與評定學習的評量（assessment of learning）之間——這是 Shepard (2000) 所辨識的評量改革的根本張力。如果支持學生學習的同一 AI 系統也為評鑑決定產出證據，利益衝突便會出現：學生和機構可能會策略性地管理其與 AI 的互動，以最佳化評鑑結果而非學習。在整合式代理系統中同時維持形成性與總結性功能的完整性，需要架構性的分離——這是一個評量設計的挑戰，ADAPT 框架的 Trust and Ethics（信任與倫理防護）組成要素（第六節）直接處理這一問題。

表演性的挑戰

在進入最後一個面向之前，必須處理一個重要的反論。台灣採用成果導向教育（outcomes-based education, OBE）的經驗，提供了一個制度表演性（institutional performativity）的實證案例——機構採用改革的語言和文件，卻未必改變改革旨在改善的底層實踐。如 Lin et al. (2021) 所觀察，許多台灣機構採行了 OBE 框架、產出了令人印象深刻的學習成果文件、製作了與能力框架對齊的課程地圖、並產出了評鑑所要求的證據物件——同時教學與評量實踐在實質上維持不變。文件是真實的；它所記載的轉型，在許多情況下卻是表面的。

沒有理由假設 AI 增強評量可以免於表演性動態。機構可能為了合規目的而採用 AI 評量工具——部署學習分析儀表板、產出 AI 處理的能力報告、並在評鑑文件中呈現這些物件——同時繼續 AI 工具名義上補充的傳統實踐。如果評鑑標準獎勵的是 AI 工具的存在而非其對學習的影響證據，則風險特別嚴峻：機構會被激勵去獲取和展示技術，而非將其有意義地整合至其評量生態中。更糟糕的是，機構可能學會最佳化 AI 生成的指標而未達到真正的學習改善——一種古德哈特定律（Goodhart’s Law）的動態，其中測量本身成為了目標。

數項設計特徵可以緩解 AI 增強評量中的表演性。首先，過程導向的評量設計：評鑑標準不應評估機構是否擁有 AI 工具，而應評估這些工具如何被使用——它們是否追蹤學習行為和過程（而非僅是產出）、其產出是否實際影響了教學調整、以及這些調整是否產出了

可測量的學習改善。其次，透過跨機構標竿比較進行外部驗證：個別機構的 AI 生成能力檔案應定期以外部測量進行驗證——標準化評量、雇主評價、由外部審查者進行的畢業專題審查——以防止機構對內部指標的操弄。第三，學生聲音機制：學生對 AI 中介評量如何影響其學習經驗的自述，提供了一種難以捏造的制衡，以對抗機構自報。第四，定期再校準：AI 評量系統應定期以學習的直接測量進行再校準，以確保其追蹤的能力對應的是真正的教育成果，而非測量過程本身的假象。第 5.4.3 節所提出的 AI 系統稽核檢核表應明確將表演性指標納為其評估標準的一部分。

4.3.7 評量者（***Assessor***）：從人類審查者到人機協作評量

最後一個面向涉及評量者的身分。現行典範將評量權威賦予人類專家——評鑑小組成員、外部審查者、學科專家——其專業判斷構成品質保證的黃金標準 (HEEACT, 2023)。代理式典範並非消除人類評量者，而是透過多代理人協作與人在迴圈中的監督（human-in-the-loop oversight），根本性地重構其角色。

在此重構模型中，AI 代理人處理證據蒐集、模式偵測與初步分析等勞力密集的工作，而人類審查者則專注於需要專業知識、脈絡知識與倫理判斷的詮釋、評價與審議任務。此種分工反映了 Zhong and Zhao (2025) 所刻畫的 AI 時代教育典範中人工智慧與人類智慧之間的必要互補性：AI 擅長規模化處理、一致性與模式識別；人類擅長意義建構、脈絡判斷與涉及價值的審議。

風險是去技能化（deskilling）。如果人類審查者越來越依從 AI 生成的分析，其自身的評量專業可能會萎縮，造成一種依賴性，破壞了系統旨在保存的人類判斷 (Bearman & Ajjawi, 2023)。第 4.5 節將詳細處理此一風險。

4.4 情境範例：一位台灣科技大學的學生

為了將七大面向紮根於具體經驗，請考慮以下情境。這個情境是刻意寫實的——設定在既有的機構結構與近程技術能力之中——而非推測性的。

韋霖是一位台灣科技大學機械工程系的三年級學生。在現行典範下，韋霖的學習將透過課程成績、畢業專題評量來測量，並在畢業數年後透過就業調查詢問其工作是否與就讀領域相關。這些資料點將與其世代同儕彙總，並報告於該系下一次的自我評鑑報告中。

在代理式典範下，韋霖的學習歷程被持續地、以多模態方式觀察。一個與大學學習管理系

統整合的 AI 代理人追蹤她與課程教材的互動，記錄的不僅是完成與否，還包括互動模式：她重新造訪哪些問題、在哪裡尋求額外資源、她的工程設計問題取徑如何跨學期演化。當韋霖提交一份計算力學作業的程式碼時，代理人分析的不僅是正確性，還包括問題解決策略——她是否系統性地分解問題、是否測試邊界條件、她的文件記錄是否反映出對其推理過程的後設認知覺察（metacognitive awareness）。此分析援引以證據為中心的設計原則 (Mislevy et al., 2003)，維持對被評量之能力宣稱及支持這些宣稱之證據的明確模型。

在她的第二學期，AI 代理人偵測到一個模式：韋霖在需要空間推理的課程中——電腦輔助設計、材料科學視覺化——表現停滯，而她的分析與計算技能則持續發展。代理人並非等待這個模式以低分呈現——或更糟，完全未被偵測——而是為韋霖的學業導師生成一則形成性警示，包括對可能原因的初步分析以及一組建議介入措施：額外的視覺化練習、與一位空間推理能力強的同學配對同儕輔導、與授課教師會面討論設計問題的替代取徑。

韋霖本人與 AI 代理人作為學習中的協作者互動。她使用代理人建構一份能力歷程檔案——一個動態的、證據豐富的發展中專業能力表徵，由她與 AI 的觀察共同策展。當代理人辨識出一項她未曾認識到的優勢——在團隊專案中整合跨領域觀點的非凡能力——她將其添加至歷程檔案，附上 AI 的佐證。當她不同意代理人對其技術寫作的評量——她認為它低估了她向非專業讀者溝通複雜概念的能力——她為此歧見加上註解，提供來自業界實習報告的反證。歷程檔案因此成為一個建設性對齊（constructive alignment）(Biggs, 1999) 的場域，不是在學習目標與評量任務之間，而是在 AI 觀察與人類自我理解之間。

在韋霖的畢業專題——為一家當地製造商的生產線設計零組件——中，一個多代理人系統跨多個面向協調評量。一個代理人評估她工程設計的技術健全性，援引領域專業知識庫。另一個代理人分析她的專案管理行為：她如何在團隊中分配任務、如何回應挫折、如何整合業界導師的回饋。第三個代理人綜合她整個軌跡的證據，將畢業專題表現置於其四年發展的脈絡中。教師評量者收到的不是一個成績建議，而是一份證據敘事——一個關於韋霖知道什麼、能做什麼、以及她作為一位工程師已成為什麼的豐富、脈絡化的敘述，由多模態證據與縱貫軌跡分析所支持。

就 HEEACT 評鑑而言，韋霖的學習證據——經匿名化並以適當的隱私保護進行彙總——貢獻於一幅持續更新的學程學習成果圖像。評鑑委員不需等待自我評鑑報告才能得知該學

程的畢業生是否發展出其課程所承諾的能力；他們可以存取顯示跨世代能力軌跡的即時儀表板、在新興缺口成為系統性問題之前辨識它們，並將有限的實地訪視時間集中於需要人類判斷的詮釋性與評價性問題上。

此情境並非科幻小說，但也不是對當前能力的準確描述。其每一個個別元素——學習分析、適性評量、能力歷程檔案、多代理人系統——今日都以某種形式存在 (Swiecki et al., 2022; Shute & Ventura, 2013; OpenAI, 2025)。代理式典範所增添的是整合：將這些元素連結為一個連貫的、持續運作的、目標導向的評量系統的能力。當前能力與此整合願景之間的落差仍然顯著。

學術誠實要求我們依可行性時程對情境中的要素進行分類：

- 近程可行（**2026-2028**）：使用大型語言模型對書面作業的自動回饋；追蹤學習管理系統中學生參與度與表現的基本學習分析儀表板；與學程學習成果連結的能力標記評量規準。這些能力今日已存在，且在全球各地的機構中以不同成熟度部署。
- 中程可能（**2028-2030**）：具備持久記憶的跨課程能力追蹤，使 AI 系統能跨學期維持學習者模型；根據累積證據調整評估策略的適性評量排序；將書面作業、專案製品與參與資料綜合為連貫能力檔案的多模態證據整合。這些能力已在原型或有限部署中被證明，但面臨重大的工程、資料基礎設施與機構整合挑戰。
- 長程推測（**2030** 年以後）：完全自主的多代理人評量生態系統，其中專門化的 AI 代理人跨畢業專題的多個面向協調評量；將學生層級軌跡資料彙總為學程與機構層級品質指標以供評鑑使用的即時機構品質儀表板；將四年縱貫資料綜合為對個別學習發展之細緻敘述的 AI 生成「證據敘事」。這些能力在技術上是可構想的，但尚未在大規模教育場域中被證明。

如上所呈現的韋霖情境代表了一個三個時程同時運作的複合願景——此一狀態在分析上有助於說明典範轉移的全面意涵，但不應被誤認為近程實施計畫。第 5.3 節所提出的分階段實施路徑，旨在依序穿越這些時程，將每個階段紮根於其前一階段所產出的實證證據。

4.5 人類判斷在新典範中的角色

代理式人工智慧所激發的熱情，必須以對其所不能為之事的清醒評估來加以節制。Bearman and Ajjawi (2023) 提醒「現在要斷言生成式 AI 是否或將會是人們所預測的典範轉移，仍為時過早」（p. 1），此一審慎態度值得重視。ADAPT 框架堅持，典範轉移並非從人類評量到 AI 評量，而是從僅有人類的評量到人機協作評量——此一區別在概念上至為關鍵，在實踐上深具後果。

半人馬模型（**The centaur model**）。在競技西洋棋中，最強大的棋手既非單純的人類也非單純的 AI，而是結合人類策略直覺與 AI 運算能力的「半人馬」團隊 (Kasparov, 2017)。評量中的半人馬模型將 AI 持續性、大規模、多模態的證據處理能力與人類的脈絡詮釋、倫理判斷與意義建構能力相結合。兩者單獨都不充分。AI 代理人可以偵測到一位學生的協作行為在一個學期間發生了變化；但需要人類評量者來詮釋這種變化究竟反映的是深化的領導技能、因個人困難導致的社交退縮，還是團隊環境中的策略性搭便車。

教師作為後設評量者。在代理式典範中，教師並非放棄評量責任，而是將其提升。教師不再花費數小時批改作業——這是 AI 日益勝任的任務——而是成為後設評量者（meta-assessors），評估 AI 生成證據的品質與 AI 生成能力宣稱的效度。此角色需要新形式的專業知識：涵蓋理解 AI 系統如何生成證據、可能編碼哪些偏見，以及其判斷在何處需要人類覆議的評量素養（assessment literacy）。Shepard (2000) 曾主張評量改革需要教師培育的相應轉型；代理式典範使此一要求變得迫切。

HEEACT 評鑑委員作為系統稽核者。評鑑委員的角色同樣轉型。他們不再評估機構的自我評鑑報告是否準確反映其實踐——此任務受限於資訊不對稱與印象管理——而是評估機構的 AI 評量系統是否產出有效、可靠、公平且透明的證據 (Pellegrino et al., 2001)。這是一個要求更高的角色，需要同時理解學科內容與 AI 系統行為，但也是一個更具影響力的角色：稽核產出證據的系統比稽核個別證據物件更有價值。

不可化約的人類領域。學習的某些面向抗拒演算法評量，不是因為 AI 尚未足夠精密，而是因為它們在構成上就是人類的。創造力——不是 AI 擅長的新穎組合生成，而是真實個人願景的表達。倫理推理——不是 AI 勝任的規則對案例的應用，而是在價值衝突且無演算法可仲裁之處的真正道德困境的導航。人際成長——同理心、文化敏感度，以及真實

人際連結能力的發展。這些學習面向並非留待日後處理的邊緣案例；它們是高等教育使命的核心，而其評量需要任何代理式系統都無法複製的人類判斷 (Pellegrino et al., 2001)。ADAPT 框架堅持人在迴圈中設計，不是對當前技術限制的讓步，而是對教育中不可化約之人類面向的原則性承諾。

4.6 從「學生成功」到「學習軌跡」：重新定義我們測量的對象

本節所描述的典範轉移不僅是方法論上的——如何測量的改變——更是本體論上的：測量什麼的改變，以及推而廣之，我們重視什麼的改變。本最後一小節論證，代理式典範最深遠的意涵在於教育品質本身的重新定義。

端點測量的貧乏。當前的「學生成功」指標——畢業率、就業率、證照考試通過率、學業成績平均——共享一個特徵：它們是捕獲目的地而未照亮旅程的端點測量（endpoint measures）。90% 的畢業率告訴我們的，對於畢業生學了什麼、如何學習、或者機構的教育實踐是否對其發展有所貢獻還是僅僅認證了先前的成就，什麼也沒說。UNESCO (2025) 的挑釁性提問——「什麼值得被測量？」——挑戰評量社群超越容易量化的端點，走向更豐富、更混亂、更具後果性的問題：學生如何成長。

學習軌跡作為新的分析單位。代理式典範使評量的基本單位從端點轉移至軌跡（trajectories）——學習者隨時間發展能力、傾向與認同的動態路徑。Zhong and Zhao (2025) 對 AI 時代教育典範轉移的分析體現了這種重新定位，主張跨時間與空間面向對齊的持續性形成性評量——追蹤跨學期、跨課程、跨學程的學習發展——必須取代定期性的總結性快照。代理式人工智慧提供了在規模上操作化此種時間對齊評量的基礎設施，維持個別學習軌跡的持久模型，不僅捕獲學生在任何給定時刻知道什麼，更捕獲其知識、技能與傾向如何發展。

過程重於產出。軌跡導向的典範將過程（process）置於產出（product）之上。學生如何處理一個不熟悉的問題？她是否系統性地分解它還是以嘗試錯誤法進行？她如何回應失敗——以堅持、策略調整還是迴避？她如何將來自同儕、教師與 AI 代理人的回饋整合至其後續工作中？這些過程導向的問題比產出導向的問題（「學生是否得到正確答案？」）更具診斷價值，也更能預測長期的專業成功 (Pellegrino et al., 2001)。代理式人工智慧持續觀察與多步驟推理的能力，使過程導向評量在前所未有的規模上變得可行。

能力軌跡與後設認知發展。超越內容知識與技術技能，代理式典範使追蹤後設認知發展（metacognitive development）成為可能——學生對自身學習過程日益增長的覺察、自我調節的能力，以及跨情境遷移學習的能力。Banihashem et al. (2025) 證明，為形成性目的最佳化的學習分析可以提供後設認知過程的豐富證據，包括自我監控、策略選擇與反思調整。當與代理式 AI 的持久記憶和適應性行為結合時，此類分析可以產出縱貫性後設認知檔案，捕獲的可能是高等教育最重要的成果：學會如何學習的能力。

本體論轉移。最終，從靜態測量到動態學習證據的轉移，蘊含著教育品質的重新定義。品質不再是機構的屬性——它們所擁有並透過自我評鑑展示的東西——而是學習經驗的屬性：動態的、關係性的、湧現的。此一本體論轉移與建構主義學習理論 (Biggs, 1999) 相契合，該理論主張學習不是知識從機構向學生的轉移，而是一個主動的意義建構過程。代理式典範首次提供了與此理論承諾相稱的評量基礎設施。

典範轉移的七個面向——時間性（temporality）、粒度（granularity）、能動性（agency）、回饋延遲（feedback latency）、證據類型（evidence type）、評量目的（assessment purpose）與評量者身分（assessor identity）——共同描述了一場既非純技術性亦非純概念性、而是兩者同時發生的轉型。代理式人工智慧提供了賦能的基礎設施；ADAPT 框架提供了分析結構；而從端點到軌跡的本體論轉移則提供了概念基礎。然而，框架與技術並不轉型教育系統；政策才能。第五節轉向的問題是：台灣的品質保證政策——具體而言是 HEEACT 的評鑑標準——如何演化以賦能、規管與治理此一典範轉移，將理論可能性轉譯為制度現實。

5. 政策與評鑑意涵

前述章節已確立，代理式人工智慧（agentic AI）具備從理論層面轉變學習成果衡量方式的能力——從定期回顧式的快照轉為持續且適應性的證據流。然而，理論潛力本身無法重塑機構實務。關鍵的中介變項是政策——具體而言，是台灣教育部（MOE）與高等教育評鑑中心基金會（HEEACT）如何選擇將代理式 AI 定位於國家品質保證（quality assurance, QA）架構之中。本節透過分析將代理式 AI 整合至台灣評鑑架構的三種政策情境、進行多面向比較評估，以及提出分階段實施路徑，以回應研究問題四（RQ4）。最為關鍵的是，本節為第四週期大學校院校務評鑑的設計提供具體建議，其設計窗口代表一個即時且影響深遠的機會。本分析援引 Bardach and Patashnik (2019) 的政策分析八步法（eightfold path），同

時扎根於台灣特定的法規脈絡、機構文化，以及來自其他國家在教育領域早期 AI 實驗中所浮現的警示教訓。為使八步法的分析結構清晰可見，本節的組織方式對應 Bardach 的各步驟如下：問題界定（第三節所辨識的結構性限制以及整合代理式 AI 的政策挑戰）；證據彙整（第二節的技術能力分析及第 5.2 節的南韓警示案例）；替代方案建構（第 5.1 節的三種政策情境）；評估標準選定（表 4 的多準則評估矩陣）；結果預測（跨可行性、公平性、成本與時程的比較分析）；面對取捨（承認沒有任何情境在所有準則上都具有主導優勢）；決策（建議從情境 B 開始的分階段途徑）；以及敘事呈現（第 5.4 節的具體第四週期建議，將分析轉化為可行動的指引）。因此，Bardach 的架構不僅是一個啟發來源，更是後續政策分析的結構性原則。

5.1 三種政策情境

將代理式 AI 整合至高等教育品質保證並非一個二元命題。依循廣泛應用於技術治理的情境規劃方法論（OECD, 2023），本小節提出三種截然不同的情境，代表從最小干擾到典範替代的連續光譜上的三個定位點。每種情境對機構準備度、監管意願及技術成熟速度各有不同的假設。

5.1.1 情境 ***A*****——**保守整合（***Conservative Integration***）

在情境 A 之下，代理式 AI 作為補充性分析工具，完全在現有 HEEACT 架構內運作。第三週期的四大項目及其相關核心指標維持不變。希望部署 AI 驅動評量工具——智慧型教學系統、自動化評分、預測性學習分析——的機構可自願為之，費用自行負擔。HEEACT 評鑑委員持續檢視傳統佐證資料：課程大綱、成績分佈、學生問卷、總整課程作品集及雇主回饋。AI 產生的證據若被提交，則視為補充文件而非主要評鑑證據。

此情境的首要優勢在於其低度監管風險。無需修法、無需修訂評鑑手冊，也無需培養新的評鑑委員能力。它尊重 Lin et al. (2021) 所辨識的台灣品質保證體系之基礎原則：大學自主與公共課責的平衡。正在實驗 AI 工具的機構得以自由進行，無需擔憂其創新將因評鑑委員不熟悉該技術而受到不利評價。

然而，保守整合（Conservative Integration）具有顯著的代價。最關鍵的是，它可能造成 Coburn (2003) 所稱的「雙軌制度」——善用 AI 提升學習成果的資源充足機構與缺乏能力或誘因跟進的機構之間，差距日益擴大。當評鑑架構未能肯認創新實務時，便隱含地貶

低這些實務，向機構領導者傳遞一個訊號：AI 投資承擔聲譽風險卻得不到評鑑回報。此外，情境 A 排除了系統層級學習的可能性：若缺乏跨機構 AI 輔助評量實務的結構化資料蒐集，HEEACT 便無法建立未來政策迭代所需的證據基礎。

5.1.2 情境 ***B***——架構演進（***Framework Evolution***）

情境 B 構想 HEEACT 評鑑架構的審慎漸進式演進，以正式承認 AI 產生的學習證據。在此情境下，HEEACT 修訂部分核心指標——特別是項目三（教學與學習）之下的指標——的描述語，以納入反映 AI 輔助評量基礎建設的新面向。例如，核心指標 3-2 目前強調「學生學習成效之直接與間接評量機制」（HEEACT, 2023, p. 38），其描述語將擴展為納入「持續性學習證據機制，包括在機構認為適當之處的 AI 中介形成性評量資料」。

標準三之下可能引入新指標，如「AI 輔助評量基礎建設」及「學習分析於學生支持之應用」。自我評鑑報告（SAR；自我評鑑報告）模版將修訂，納入關於 AI 評量的選填附件，要求部署此類工具的機構記錄其資料治理、演算法透明度措施及偏誤緩減方案。關鍵的是，此情境維持現有評鑑週期結構——定期自我評鑑加實地訪評——但豐富了評鑑委員可資運用的證據選項。

情境 B 的優勢在於其兼顧創新與延續性的平衡。台灣的評鑑文化歷經三個週期、逾十五年的淬鍊而趨於成熟；機構與評鑑委員已發展出對品質證據、訪評程序及改進導向對話的共同理解（Lin et al., 2020）。情境 B 善用而非揚棄此一累積的品質保證文化。它也與若干同儕 QA 機構所浮現的做法一致：歐洲品質保證準則（ESG）的修訂討論已強調將數位評量納入既有品質架構，而非建立平行體系（ENQA, 2015）。

情境 B 的主要限制在於速度。評鑑標準的漸進修訂通常需要多年的諮詢過程，涉及教育部、HEEACT 職員、機構代表、學生團體及業界利害關係人。若 AI 輔助評量工具依當前軌跡快速演進（Gartner, 2025），架構演進可能陷入永恆的落後——標準總是落後技術一個世代。此外，情境 B 需要大量的 HEEACT 職員能力建構；評鑑委員必須具備足夠的技術素養，以有意義地評估 AI 評量系統，而非僅勾選合規欄位。

5.1.3 情境 ***C***——典範替代（***Paradigm Replacement***）

情境 C 代表最為激進的轉向：以即時品質監測的全新持續保證模式，取代定期 SAR 加實地訪評的結構。在此情境下，機構維護 AI 中介的證據儀表板，持續向 HEEACT 傳輸學習

成果資料。代理式 AI 系統即時產生、策展並呈現學生學習的證據——非靜態表格於每六年提交一次的文件中，而是 HEEACT 可隨時存取的活資料生態系統。實地訪評若保留，亦轉型為由持續資料流中的異常所觸發的針對性「深度審查」，而非例行性排程。

在情境 C 之下，HEEACT 的角色發生根本性轉變：從定期評鑑者轉為持續品質夥伴。HEEACT 不再每六年組建評鑑團隊，而是與機構維持持續的分析夥伴關係，提供形成性回饋、標竿比較資料及早期預警訊號。此願景與 Temper et al. (2025) 在 HEAT-AI 架構中所描述的「嵌入式品質智慧」一致——AI 系統不僅衡量品質，更主動參與品質改善。

情境 C 的吸引力在於其概念上的優雅性：它完全實現第二節理論架構中所闡述的從「學習的評量」（assessment-of-learning）到「作為學習的評量」（assessment-as-learning）的典範轉移。它消除了長期以來被認定為定期品質保證體系結構性弱點的評量時間差問題——資料蒐集與評鑑判斷之間的數月或數年落差（Ewell, 2009）。

然而，障礙是巨大的。僅基礎建設成本——標準化資料架構、每所高等教育機構與 HEEACT 之間的安全 API 連線、AI 系統維護——便已相當可觀。公平性（equity）關切至為重要：台灣高等教育體系涵蓋 152 所機構，範圍從擁有健全資訊部門的研究型大學到數位基礎建設極為有限的小型技術學院（MOE, 2024）。持續保證模式可能恰恰排除最需要品質支持的機構。國際認可呈現另一挑戰；全球尚無主要 QA 架構完全實作持續保證，意味著台灣將在缺乏既有標竿的情況下開拓前行。最後，南韓在 AI 驅動教育轉型方面的經驗——如下所述——提供了一個令人清醒的提醒：技術部署的速度並不保證教育成果的品質。

5.2 情境比較分析

為系統性評估三種情境，表 4 呈現多準則評估矩陣。評估標準反映對台灣政策脈絡最為攸關的面向：實施可行性、跨不同機構類型的公平性影響、品質保證完整性維持、國際認可相容性、預估成本、實施時程，以及機構準備度需求。

表 **4**

代理式 *AI* 整合情境之比較評估矩陣

評估標準	情境 A：保守整合	情境 B：架構演進	情境 C：典範替代
可行性	高——無需法規變更	中等——需修訂 HEEACT 手冊	低——需新立法及基礎建設
公平性影響	負面——擴大資源充足與資源不足機構之間的差距	中等——選填式指標降低準備度不足機構的壓力	高風險——可能排除缺乏數位基礎建設的機構
品質保證完整性	維持——現有標準不變	維持並增強——新指標增加嚴謹性	不確定——未經驗證的模式，無國際先例
國際認可	高——與現行華盛頓協定、ESG 一致	高——與 ESG 修訂走向一致	初期偏低——尚無同儕 QA 機構採用此模式
預估成本	極低（僅機構層級）	中等（HEEACT 能力建構 + 手冊修訂）	極高（國家基礎建設 + 持續維護）
時程	立即可行	標準修訂需 2-4 年	全面實施需 5-10 年以上
機構準備度需求	低——自願採用	中等——機構須記錄 AI 實務	極高——所有機構須維護即時資料系統
非預期後果風險	低，但有停滯風險	中等——取決於指標設計品質	高——南韓警示案例之類比

比較分析顯示，沒有任何單一情境在所有標準上都具有主導優勢。情境 A 將風險最小化，但機會成本最大。情境 C 將轉型潛力最大化，但近期成本與風險過高。情境 B 佔據務實的中間地帶，但需要審慎設計以避免過於保守（趨向 A）或過於躁進（過度延伸至 C）。這一模式——溫和選項浮現為最可行的基礎——與 Bardach and Patashnik (2019) 的觀察一致，即有效的政策分析通常收斂於透過漸進步驟實現激進目標的混合途徑。

南韓的警示案例

任何關於快速部署 AI 於教育領域的討論，都必須正視南韓的經驗。2024 年，南韓教育部宣布一項雄心勃勃的計畫，將在 2025 年前於所有學校部署 AI 驅動的數位教科書，將此倡議定位為教育現代化的基石。推動速度極快——而反彈同樣劇烈。家長組織抗議，關切螢幕時間增加及人際互動減少。教師反映 AI 教科書對多元學習者需求的校準不佳，且許多學校的支援基礎建設不足。到 2025 年中，政府被迫大幅回調，延後全面實施並縮減 AI 整合的範圍（Rest of World, 2025）。

南韓案例揭示三項直接適用於台灣審議的教訓。第一，技術就緒不等於利害關係人就緒；即使 AI 系統按設計運作，教師、學生和家長可能尚未準備好面對其所帶來的教學轉變。第二，近用公平是前提而非附帶考量；在未確保所有機構都能支持 AI 工具的情況下部署，將製造侵蝕公眾信任的明顯不公。第三，伴隨真實評估的分階段試辦——而非為預定推動方案背書的形式性試辦——是不可或缺的。這些教訓強烈反對將情境 C 作為近期策略，並強化了第 5.3 節所詳述之分階段途徑的論據。

5.3 建議之分階段途徑

基於情境分析與南韓的警示性證據，本文建議一條三階段實施路徑：始於保守試辦，基於實證過渡至架構演進（Framework Evolution），並為已展現準備度的機構選擇性引入典範轉型要素。

第一階段：結構化試辦（**2026-2028**）。在第一階段，教育部與 HEEACT 聯合遴選 10-15 所涵蓋多元機構類型的高等教育機構——研究型大學、綜合大學及科技大學——參與 AI 輔助評量的結構化試辦。參與機構獲得適度的經費支持與技術指導，以在選定系所實施代理式 AI 工具。HEEACT 制定記錄 AI 產生學習證據的暫行指引，評鑑委員對 AI 評量實務進行補充性審查（非正式評鑑）。關鍵的是，第一階段包含嚴謹的評估機制：由獨立研究人員評估 AI 輔助評量在各試辦場域的效度（validity）、信度（reliability）、公平性及教學影響。台灣於 2025 年 12 月通過的人工智慧基本法，以其七大治理原則——包括人類自主、隱私保護、透明及公平（Legislative Yuan, 2025）——為確保試辦實施符合國家 AI 治理標準提供了法律架構。

第二階段：架構演進（**Framework Evolution**）（**2028-2030**）。第二階段僅在第一階段評估

資料證明 AI 輔助評量能產生有效、公平且具教學價值的學習證據後方啟動。基於試辦發現，HEEACT 啟動項目三及項目四之下核心指標描述語的正式修訂，納入 AI 中介評量證據的新語彙。SAR 模版更新以納入情境 B 所述之 AI 評量附件。HEEACT 啟動評鑑委員能力建構計畫，包括赴同儕 QA 機構的國際標竿參訪（見第 5.5 節）及 AI 系統稽核專業訓練模組。第二階段與第四週期校務評鑑的預期啟動時間窗口一致，為修訂後的標準創造自然的整合點。

第三階段：選擇性轉型（**2030** 年以後）。對於在第一及第二階段已展現成熟 AI 評量基礎建設的自辦評鑑機構，第三階段選擇性引入情境 C 的要素：持續品質監測儀表板、即時證據流，以及機構與品質保證機構之間更具流動性的關係。關鍵的是，第三階段維持傳統評鑑路徑作為完全正當的替代方案。沒有任何機構被強制採用持續保證模式；它而是成為數位成熟度足以支撐的機構的一個選項。此雙軌模式確保系統不會複製南韓回調案例中所觀察到的公平性失敗（Rest of World, 2025）。

5.3.1 實施的政治經濟學

前述分階段途徑預設了利害關係人之間一定程度的協調，而台灣高等教育治理的政治經濟學並不保證這一點。關鍵行動者——教育部、HEEACT 與大學——之間的關係涉及重疊但各異的權限、誘因與限制，必須明確加以處理。

此一轉型的政治經濟學不應被低估。HEEACT 作為獨立財團法人的定位提供了運作彈性，但限制了其強制推動 AI 採用的權限；教育部掌控經費但可能優先考量其他政策目標；而大學面對有限資源的競爭性需求。具體而言，HEEACT 發展評鑑標準並管理評鑑流程，但教育部保有對架構整體設計及評鑑結果之法規後果的核准權限。大學實施評量實務，但自主程度各異：自辦評鑑機構享有相當的程序彈性，而接受標準評鑑的機構則在較緊密的法規約束下運作。

經費配置是一個特別棘手的協調挑戰。誰來資助即使是適度試辦所需的 AI 基礎建設？三種潛在資金來源各有局限。第一，教育部可透過高等教育深耕計畫（Higher Education Sprout Project）配置資源，該計畫已將科技增強教學品質列為資助面向之一；然而，深耕計畫經費具競爭性且有時間限制，產生永續性風險。第二，機構可從自有預算資助 AI 評量基礎建設；然而，最需要改善評量的機構恰恰是預算最受限的（見下文第 5.3.2 節）。第

三，與教育科技供應商的公私合作可提供基礎建設；然而，此類夥伴關係引發供應商依賴及資料治理的關切，必須審慎管理（見下文第 5.3.3 節）。

除教育部與 HEEACT 之外，多個額外的利害關係人有需要協調的利益：台灣 AI 學院聯盟（TAICA），可提供共享基礎建設與技術專長；各大學校長與教務長，必須擁護或至少容許 AI 評量整合；教師工會與學術治理機構，其支持對任何評量實務的變革皆不可或缺；學生會，其成員是評量轉型的最終對象；以及科技供應商，其商業誘因可能與教育價值不一致。將這些多元利害關係人圍繞共同的實施願景加以整合，需要 HEEACT 在先前評鑑週期轉換中已成功管理的那種審慎、包容的諮詢過程——但在壓縮的時程之下，鑑於技術變革的步伐。

5.3.2 資源悖論

任何 AI 輔助評量提案的底層都存在一個根本的資源悖論（resource paradox）：對少子女化危機最為脆弱的機構——招生下降、預算萎縮的偏鄉私立小型大學——恰恰是最缺乏投資代理式 AI 部署所需之 AI 基礎建設、技術專才及資料科學人力的機構。若無刻意的政策介入，AI 輔助評量可能加深而非縮小資源充足與資源不足機構之間的品質鴻溝。

數項公平性機制可因應此悖論。第一，TAICA——擁有 55 所成員大學——可作為共享 AI 評量基礎建設的載體，提供各機構無需獨立開發或維護的雲端評量工具。第二，教育部可設立專項經費用於 AI 評量準備，以現有數位基礎建設補助為模型，其資格標準優先考量資源不足的機構。第三，透過公共經費研究所開發的開源評量工具可降低進入門檻。第四，區域性 AI 評量中心——或許以資源豐富的機構為核心，但服務同一地理區域內的小型機構群——可提供任何單一小型機構無法獨立維持的技術支援、培訓及共享基礎建設。這些公平性機制並非後續政策建議的可選附加項；它們是公平實施的前提條件。

5.3.3 AI 供應商動態

代理式 AI 評量系統最可能的提供者是商業科技公司，其利潤動機可能與教育價值不一致。三項具體風險值得關注。第一，供應商鎖定（vendor lock-in）：若機構採用專屬 AI 評量平台，可能對單一供應商在系統維護、更新及資料存取方面形成依賴，限制機構自主並造成長期成本上升。第二，資料提取：商業供應商可能尋求將學生學習資料用於產品開發、廣告或超出教育關係的次級商業用途。第三，誘因錯位：供應商被激勵展示令人印象深刻的

能力與快速採用，這可能與本文所倡議的審慎、循證、以公平為中心的途徑相衝突。為緩減這些風險，HEEACT 的「AI 評量標準」（於第 6.4 節提議）應包含資料所有權的明確條款、防止供應商鎖定的互通性要求，以及學生資料不得用於教育關係以外目的的契約保證。開放標準（xAPI、cmi5、Open Badges）應為任何部署於認可課程脈絡中的 AI 評量系統之必要條件，確保學生能力紀錄不論產生該紀錄的供應商平台為何，皆可攜帶轉移。

5.4 對第四週期校務評鑑之意涵

第三週期大學校院校務評鑑於 2025 學年結束；第四週期的設計代表將台灣品質保證架構定位於代理式 AI 時代的最即時且最具影響力的機會。本小節針對第四週期設計的五個面向提供具體、可行的建議：核心指標修訂、SAR 模版更新、實地訪評程序調整、自辦評鑑機構的角色，以及評鑑委員能力建構。

5.4.1 修訂項目三核心指標

第三週期的項目三「教學與學習」包含四項核心指標（3-1 至 3-4），分別涵蓋課程設計、評量機制、學習支持及教學品質改善（HEEACT, 2023）。表 5 針對其中三項指標提出具體修訂建議，以容納 AI 產生的學習證據，同時保留該項目既有的評鑑邏輯。

表 **5**

項目三之第四週期核心指標修訂建議

核心指標	現行描述語（第三週期）	建議新增面向（第四週期）	理據
3-1：課程設計與校務目標	檢視機構使命、系所學習成效與課程結構之對應	新增：「數位評量基礎建設準備度——機構展現蒐集、管理及分析數位學習證據的能力，包括部署場域之 AI 中介評量資料」	確保發展 AI 評量工具的機構具備基礎建設；不強制採用 AI，但承認其為課程實施的合理面向

核心指標	現行描述語（第三週期）	建議新增面向（第四週期）	理據
3-2：學生學習成效之評量	評估直接與間接評量機制；檢視評量結果用於持續改善之情形	新增：「持續性學習證據機制——機構得透過持續、科技中介的證據流（如學習分析、AI 形成性評量資料、能力追蹤系統），與傳統定期評量並行或作為補充，展現學習成果之達成」	擴展證據選項而不取代傳統評量；與 UNESCO (2023) 多元評量模式指引一致
3-3：學生支持與學習資源	檢視課輔、諮商、學習資源及對多元學習者之支持	新增：「AI 輔助學習分析用於學生支持與早期介入——部署預測分析或 AI 驅動早期預警系統的機構，評鑑委員檢視其有效性、公平性及與人工諮詢的整合」	肯認學習分析儀表板日益普及之使用，同時堅持人在迴路中（human-in-the-loop）的支持結構；透過檢視 AI 工具是否服務所有學生族群以回應公平性

數項設計原則支撐這些提案。第一，每項新增均以「附加面向」而非替代方式框架，與尚未採用 AI 工具的機構保持向後相容。第二，用語一致採用條件式措辭——「在部署場域」、「機構得展現」——以避免強制要求特定技術。第三，每項建議面向皆明確提及公平性與人類監督，反映人工智慧基本法之公平與人類自主原則（Legislative Yuan, 2025）。第四，提案經校準為可評鑑的：評鑑委員可評估機構的數位基礎建設是否足夠、持續性證據機制是否產生有效資料，以及 AI 驅動的學生支持是否觸及所有學習者群體。

5.4.2 SAR 模版更新

自我評鑑報告是機構向 HEEACT 呈現品質證據的主要書面載體。第四週期 SAR 模版應納入兩項新元素：

AI 評量附件。部署 AI 驅動評量工具的機構將填寫一份選填附件，記錄：(a) 使用中的特定 AI 工具、其目的及部署範圍；(b) 資料治理方案，包括資料來源、儲存、存取控制及保留政策；(c) 演算法透明度措施——機構如何確保 AI 產生的判斷（如自動評分、學習軌跡預測）對教師與學生而言是可理解的；(d) 偏誤測試結果——AI 工具已就性別、社經地位、身心障礙及原住民身分等學生子群體的差異性表現進行評估的證據；以及 (e) 學生同意與溝通方案——學生如何被告知 AI 在其評量中的角色及存在哪些退出機制。此附件援引 Temper et al. (2025) 所提出的架構，並與台灣人工智慧基本法之透明性及隱私要求一致。

學習證據儀表板選項。機構得以即時學習證據儀表板取代或補充傳統靜態資料表（如成績分佈、畢業率、雇主調查結果），提供予評鑑委員存取。此類儀表板將呈現縱貫式學習成果資料，包括形成性評量趨勢、能力達成軌跡及學習分析摘要。為確保可比較性，HEEACT 將公布最低顯示標準，明定必要資料欄位、視覺化格式及資料時效性要求。此選項肯認在資料豐富的環境中，於訪評前數月提交之文件中的靜態表格已逐漸成為過時的證據呈現方式（OECD, 2023）。

5.4.3 實地訪評程序調整

實地訪評仍是台灣評鑑流程的基石，提供評鑑委員僅靠書面證據無法獲得的脈絡理解（HEEACT, 2023）。三項調整將為實地訪評程序迎向代理式 AI 時代做好準備。

第一，評鑑委員之能力要求應擴展，納入評估 AI 評量系統品質與完整性的能力。這並不意味每位評鑑委員都必須是機器學習專家；而是每個評鑑團隊至少應有一位成員具備充足的技術素養，能以知情的懷疑態度檢視機構的 AI 工具——例如理解一個經過驗證的適性評量引擎與一個缺乏心理計量基礎、表面上令人印象深刻的聊天機器人之間的差別。

第二，HEEACT 應發展供實地訪評使用的 AI 系統稽核檢核表。援引新加坡的 AI Verify 工具包（IMDA, 2020）及 TEQSA 的學習分析指引（TEQSA, 2024），該檢核表將涵蓋：資料來源及其出處；模型訓練程序及驗證證據；演算法透明度與可解釋性規定；偏誤測試方法與結果；系統可靠性及故障模式處理方案；學生同意文件；以及教師對 AI 評量監督的

治理結構。此檢核表具有雙重目的：為評鑑委員提供結構化的評估工具，並向機構傳達 HEEACT 認為不可或缺的 AI 治理特定面向。

第三，實地訪評的訪談程序應擴展，納入關於 AI 治理、教師 AI 素養及學生 AI 中介評量經驗的問題。訪談問題示例可包括：「學校的評量委員會如何監督 AI 驅動的評量工具？」（對行政主管）；「您接受過哪些解讀 AI 產生之學習分析的培訓，以及這些分析如何影響您的教學？」（對教師）；以及「您是如何被告知 AI 在您課程評量中的角色的，您是否認為該過程透明且公平？」（對學生）。這些問題確保 AI 整合的人文面向——治理、素養、信任與經驗——獲得與技術面向相當的評鑑關注。

5.4.4 自辦評鑑機構作為先行者

台灣的品質保證體系包含一類自辦評鑑機構——已取得自行辦理評鑑程序資格、由 HEEACT 進行後設評鑑（meta-evaluation）的大學。這些機構通常是擁有成熟品質保證文化的研究型大學，具備兩項特性使其成為 AI 輔助評鑑實務的天然先行者。第一，它們擁有更大的程序彈性；其評鑑程序無需在所有細節上遵循標準 HEEACT 模版。第二，它們通常掌握更充裕的財務與技術資源，降低了制約資源不足機構的公平性障礙。

HEEACT 應發展「AI 就緒自辦評鑑指引」（AI-Ready Self-Accreditation Guidelines），使自辦評鑑機構在建議實施路徑之第二階段期間，得以擔任情境 B 實務的試辦場域。這些指引應明定：AI 評量工具文件記錄的最低要求；效度與公平性測試的預期證據；AI 監督的治理結構；以及允許 HEEACT 跨自辦評鑑機構彙總學習成果的報告方案。自辦評鑑機構的後設評鑑流程屆時將納入 AI 整合成熟度之評估作為補充面向，產生推動更廣泛架構演進所需的機構證據基礎。

5.4.5 評鑑委員培訓與能力建構

第四週期準備中最為關鍵——也最常被低估——的要素，或許是評鑑委員的能力建構。台灣的評鑑委員主要是志願貢獻專業的資深學者與行政主管；鮮少具備人工智慧、學習分析或教育資料科學的背景。若缺乏針對性的能力建構，即使設計良好的指標與程序也將流於表面執行。

HEEACT 應建立三個構成部分的評鑑委員發展計畫。第一個構成部分是 AI 評量的基礎工作坊，涵蓋核心概念（機器學習、自然語言處理、學習分析）、高等教育中的常見應用，

以及評鑑委員在檢視 AI 系統時應提出的關鍵問題。這些工作坊無需培養技術專家；而應培養 Selwyn (2019) 所稱的「批判性數位素養」（critical digital literacy）——能對技術宣稱提出有穿透力的問題，而不被技術術語所震懾。

第二個構成部分是國際標竿。HEEACT 應組織赴 AI 治理最為先進之同儕 QA 機構的參訪學習——特別是新加坡私立教育委員會（CPE），受益於國家級 AI 治理模式架構與 AI Verify 測試工具包（IMDA, 2020）；澳洲高等教育品質與標準署（TEQSA），已就學習分析及 AI 於高等教育之應用公布領先指引（TEQSA, 2024）；以及參與 ENQA 數位品質保證工作小組的若干歐洲機構。這些參訪將使 HEEACT 職員與評鑑委員候選人接觸國際最佳實務，並協助台灣避免重新發明同儕機構已開發的解決方案。

第三個構成部分是創設新的評鑑委員專業類別：「數位評量專家」。持有此一資格認定的個人須已展現評估 AI 評量系統的能力，並將被指派至訪評具有重要 AI 部署之機構的評鑑團隊。隨著時間推移，當 AI 整合更為普及，與此專業相關的技能將擴散至一般評鑑委員群體；然而在近期，一個專責的專家角色確保至少有一位團隊成員能對 AI 系統進行具技術知情性的評估。

5.5 國際比較參照

台灣並非孤立地面對 AI 輔助品質保證的挑戰。對同儕 QA 機構回應 AI 的簡要調查，既提供標竿也提供警示性的參照點。

歐洲品質保證準則（ESG），最近一次於 2015 年修訂，目前正進行明確考量高等教育數位轉型的審查討論。ENQA 成員機構已發表關於品質保證中 AI 的工作文件，普遍傾向與情境 B 一致的演進式途徑——將 AI 考量整合至既有標準而非建立平行架構（ENQA, 2015）。然而 ESG 的修訂過程眾所周知地緩慢；任何更新標準最早也不太可能在 2027 年前出現，顯示台灣若採取更果決的行動，可以自我定位為區域領導者。

澳洲高等教育品質與標準署（TEQSA）是全球最為積極主動的 QA 機構之一。TEQSA 於 2024 年發布關於人工智慧在高等教育之指引，同時涵蓋教學應用與品質保證意涵。值得注意的是，TEQSA 強調提供者（provider）有責任確保 AI 中介評量的完整性——此一框架與為台灣 SAR 模版所提議的 AI 評量附件一致（TEQSA, 2024）。

英國品質保證局（QAA）已發表多篇關於 AI 與品質的立場文件，包括生成式 AI 時代之學

術誠信指引。QAA 的途徑強調適應性——鼓勵機構在國家品質期待架構內發展本校的 AI 政策——而非規定特定的技術解決方案（QAA, 2023）。

在美國，高等教育認證委員會（CHEA）已成立 AI 與認證工作小組，反映對區域認證機構必須發展 AI 相關能力之日增認知。然而，美國認證的分散特性——多個區域及專業認證機構各自獨立運作——已減緩統一指引之發展。

新加坡作為台灣的比較對象特別值得關注。資通訊媒體發展局（IMDA）的 AI 治理模式架構（Model AI Governance Framework）現已出版第二版，為跨部門——包括教育——的負責任 AI 部署提供國家級結構。AI Verify 工具包——一個開源測試架構，允許組織依據治理原則評估其 AI 系統——恰恰是 HEEACT 可以改造用於評鑑目的的實務工具（IMDA, 2020）。新加坡教育科技主計畫 2030（EdTech Masterplan 2030）闡述了以證據與公平為基礎的科技增強教育願景，提供台灣教育部在制定自身 AI 於教育路線圖時可參照的策略規劃模式（MOE Singapore, 2023）。

綜合而言，國際面貌揭示一個收斂的模式：沒有主要 QA 機構採用類似情境 C 典範替代（Paradigm Replacement）的做法，但大多數正積極超越情境 A 的保守立場，朝向某種版本的情境 B 架構演進（Framework Evolution）。台灣所建議的分階段途徑因此與國際趨勢一致，同時將國家定位為亞太區域的領先者——尤其考量 HEEACT 作為亞洲最成熟 QA 機構之一的既有聲譽及其在 INQAAHE 與 APQN 中的活躍角色（Lin et al., 2021）。

本節所呈現的政策與評鑑建議建立在一個基本假設之上：代理式 AI 整合至學習成果衡量，能以倫理、公平且透明的方式加以治理。此假設遠非不言自明。下一節深入檢視 AI 輔助評量的倫理面向，探問演算法效率與人性尊嚴之間、資料豐富性與隱私之間，以及創新與教育中人類判斷不可化約之價值之間的張力。

6. 倫理考量與風險治理

代理式人工智慧在學生成功衡量中的部署，引入了與傳統教育科技截然不同的倫理挑戰。與被動式分析儀表板或規則導向的早期預警系統不同，代理式 AI 以自主決策權限、持久記憶及迭代推理能力運作，從根本上改變了機構、教育者與學生之間的權力動態。隨著台灣朝向在高等教育品質保證基礎建設中整合此類系統邁進，嚴謹的倫理分析不僅是值得做的——而是負責任創新的先決條件。本節將 Beauchamp and Childress (2019) 的原則主義

生命倫理架構（principlist bioethics framework）應用於教育 AI 脈絡，建構全面性的風險矩陣，辨識代理式架構特有的風險，並提出一個校準於台灣法律、文化與機構場域的三層治理架構（three-tier governance framework）。

6.1 四原則倫理分析

原則主義架構（principlist framework）最初為生物醫學倫理而發展，近年來日益被技術倫理學術研究所採用，作為評估複雜社會技術系統的結構化途徑（Floridi et al., 2018）。其四項原則——自主（autonomy）、行善（beneficence）、不傷害（non-maleficence）及公義（justice）——提供一個系統性的透鏡，透過它檢視代理式 AI 在評量中的倫理意涵。

6.1.1 自主（***Autonomy***）

自主（autonomy）原則要求個人對影響其生活的決定保有有意義的控制。在代理式 AI 用於學生成功衡量的脈絡中，自主性關切沿三個相互關聯的面向展現：知情同意（informed consent）、退出權及資料主權（data sovereignty）。

持續監測學生的學習行為——登入模式、作業繳交時間、討論區參與及能力展現軌跡——引發關於同意的根本問題。Slade and Prinsloo (2013) 論證，學習分析創造了一種不對稱的權力關係，機構蒐集並分析可能未充分理解或未有意義地同意監控範圍的學生之資料。當代理式 AI 疊加於這些分析之上時，不對稱性加深：學生不再僅被觀察，而是成為 AI 代理關於其能力狀態、介入路徑甚至資格就緒度的自主決定之對象。

退出權呈現一個特別棘手的挑戰。若機構將代理式 AI 嵌入其核心評量架構，退出可能實際上意味著退出教育經驗本身——一種削弱真正同意的強制結構（Zuboff, 2019）。台灣的個人資料保護法為資料主體權利提供了法律基礎，包括請求停止資料蒐集與處理的權利。然而，該法之條文主要針對商業資料實務所擬定，其對教育脈絡——資料蒐集與教學過程交織——的適用性在法律上尚未受到檢驗。

資料所有權與可攜性構成第三個自主性關切。當代理式 AI 建構橫跨學生整個大學生涯的縱貫式能力檔案時，誰擁有該檔案？學生、機構，還是 AI 系統開發者？台灣於 2025 年 12 月通過的人工智慧基本法闡述了透明性與人類能動性原則，但尚未解決教育脈絡中特定的資料所有權問題。若缺乏明確的資料可攜性標準，學生可能發現其能力紀錄被鎖定於專屬系統中，在轉學或進入職場時無法轉移其學習歷程。

6.1.2 行善（***Beneficence***）

行善（beneficence）原則要求介入措施能產生可證明的效益以證成其成本與風險。AI 增強學習的證據基礎雖具前景，但在不同脈絡與時間尺度上的分佈仍不均。Kestin et al. (2025) 在哈佛大學進行的隨機控制試驗中發現，使用 AI 教學系統的學生所達成的學習增益約為傳統主動學習環境學生的兩倍——這一引人注目的結果已吸引相當關注。然而，此研究涉及一個特定的、資源充裕的脈絡（一所菁英研究型大學的物理學入門課程），且採用狹義的成果量測指標（立即的教學後測表現）。此類增益是否能持續、能否遷移至其他學科，或能否在許多台灣高等教育機構所典型的資源受限環境中複製，仍是有待實證的開放問題。

更廣泛而言，歸因於代理式 AI 在評量中的效益——個人化回饋、對高風險學生的早期介入、持續能力追蹤及適應性學習路徑優化——在很大程度上是理論性的，或僅在短期試辦研究中被證明（Temper et al., 2025）。UNESCO 的《生成式 AI 在教育及研究中的指引》告誡不應將 AI 的潛在效益與已證明的效益混為一談，指出 AI 在教育中的證據基礎對許多所聲稱的應用而言「仍在發展中且大致不具結論性」（UNESCO, 2023, p. 18）。因此，負責任地應用行善（beneficence）原則要求機構將代理式 AI 視為需要持續評估的實驗性介入，而非可大規模部署的已證實解決方案。

6.1.3 不傷害（***Non-maleficence***）

避免傷害的義務在代理式 AI 的脈絡中尤為迫切，因為潛在傷害在個人與系統層面皆可運作。四類傷害值得關注。

第一，監控正常化（surveillance normalization）。持續的行為監測，即使被框架為支持性的，仍有風險創造出學生內化對持續觀察之期待的教育環境。Zuboff (2019) 記錄了監控資本主義如何在商業脈絡中使行為資料的提取正常化；將類似實務延伸至教育——權力不對稱已然明顯的領域——威脅侵蝕真正學習所不可或缺的智識自由與探索性冒險。

第二，演算法偏誤（algorithmic bias）。Baker and Hawn (2022) 在對教育 AI 系統演算法偏誤的全面性回顧中，記錄了對少數族群學生的系統性低估（under-prediction），包括非裔、拉丁裔及低收入學習者。這些偏誤並非偶發而是結構性的，源自編碼了歷史性不平等的訓練資料及偏好與主流文化規範相關行為模式的特徵工程選擇。在台灣的脈絡中，原住民、新

住民子女及社經弱勢背景學生存在類似的風險——這些群體的學習行為可能偏離 AI 系統（主要以多數族群資料訓練）所辨識為學業成功指標的模式。

第三，教學去專業化（de-professionalization of teaching）。當代理式 AI 承擔原先屬於教師專業判斷範疇的評量決定責任時，教育者的專業認同與專業能力可能受到削弱。此關切並非假設性的：關於醫學中 AI 輔助臨床決策的研究已記錄了自動化如何隨時間侵蝕實務工作者的診斷技能，一種被稱為「因廢退而技能退化」（skill degradation through disuse）的現象（Parasuraman et al., 2000）。在台灣，教師在課程設計與評量方面的自主是大學治理傳統的基石，去專業化風險具有特殊的文化共鳴。

第四，AI 幻覺與錯誤傳播（AI hallucination and error propagation）。代理式 AI 系統，特別是採用大型語言模型進行回饋生成者，容易產生自信但不正確的輸出。關於 LLM 在教育脈絡中可靠性的研究已證明，輕微的輸入錯誤或模糊提示可將評量準確度降低 30% 或更多（Chinta et al., 2024）。當此類錯誤發生在代理式架構內——AI 輸出餵入後續的自主決策——錯誤複合的潛力是巨大的。

6.1.4 公義（*Justice*）

公義（justice）原則要求效益與負擔的公平分配。在台灣的高等教育體系中，國立與私立大學之間、都市與偏鄉機構之間、資源充裕的研究型大學與較小型教學導向機構之間的結構性不平等，創造了代理式 AI 的效益可能不成比例地累積於本已處優勢機構的條件。

部署代理式 AI 所需的資源——運算基礎建設、技術專才、資料科學人力及持續的系統維護——是可觀的。若無刻意的政策介入，機構間的數位落差可能擴大而非縮小，產生菁英機構運用精密 AI 驅動評量、而資源不足機構仰賴人工流程的雙軌品質保證體系。此關切並非臆測性的：Gandara et al. (2024) 分析美國高等教育中的預測分析系統後發現，演算模型對非裔學生產生 19% 的偽陰性（false negatives），對拉丁裔學生為 21%——意味著這些群體中近五分之一本會成功的學生被錯誤標記為高風險。對台灣邊緣化族群——原住民、新住民子女、身心障礙學生及經濟弱勢家庭學生——的意涵是重大的。若主要以多數群體學生資料訓練的代理式 AI 系統在未經嚴謹公平性測試的情況下被部署，它們可能系統性地錯誤描述恰恰最需要機構支持的學生之能力與潛力。

6.2 風險矩陣

表 6 將倫理分析綜合為一個結構化的風險矩陣，可作為機構決策者的實務治理工具。

表 **6**

代理式 *AI* 部署於學生成功衡量之風險矩陣

風險類別	可能性	嚴重性	緩減策略
評量中的演算法偏誤	高	高	定期進行偏誤稽核，含分群人口統計分析；多元且具代表性的訓練資料；部署前強制公平性影響評估
學生監控正常化	高	中	向學生溝通明確的資料治理政策；具真正退出機制的知情同意架構；目的限制原則
教師去專業化	中	高	所有重大評量決定的人在迴路中（human-in-the-loop）要求；教師 AI 素養計畫；將 AI 定位為增強而非取代的專業發展
資料外洩或隱私侵害	中	極高	隱私設計（privacy-by-design）架構；完全遵循個人資料保護法（個資法）；資料最小化；靜態與傳輸加密；定期滲透測試

風險類別	可能性	嚴重性	緩減策略
數位落差擴大	高	高	教育部專項基礎建設經費用於資源不足機構；透過校際聯盟之共享 AI 基礎建設；以 TAICA 架構為模型的跨機構資源匯集
過度依賴 AI 判斷	中	高	重大決定的強制人工審查門檻；具人工裁決保障的學生申訴機制；AI 與教師評量的定期校準
AI 評量學歷之國際不認可	低	中	與 ESG（歐洲高等教育區品質保證準則）原則一致；專業課程符合華盛頓協定；運用 HEEACT 之 IN-QAAHE 正式會員地位
AI 產生之評量遊戲化	中	中	與外部指標的持續驗證；多模態證據要求；抗衡指標優化的過程導向評量設計

6.3 代理式 **AI** 特有之風險

本分析的一項關鍵貢獻是區分所有教育 AI 應用共通的倫理風險與代理式 AI 架構特有的風險。既有關於教育 AI 倫理的學術研究大多處理適用於任何處理學生資料之運算系統的一

般性關切——隱私、偏誤、透明度。然而，以下所列舉的風險特別源自定義代理式 AI 的自主性、持久性及多代理特性。

自主性放大（**Autonomy amplification**）。與呈現建議供人類決策者參考的被動式 AI 工具不同，代理式 AI 以被授權的評量決策權限運作。當 AI 代理自主判定學生未達能力門檻——觸發課程不及格、補救要求或延遲畢業等後果——責任歸屬問題變得真正新穎。傳統的責任結構假設存在一個推理可被質疑、挑戰與推翻的人類決策者。代理式 AI 瓦解了這一假設。機構部署系統；開發者設計架構；AI 代理執行決定。責任被擴散至一連串行動者之間，其中任何人都可能對產生特定結果的具體推理過程缺乏完整的可見性。此「責任缺口」（accountability gap; Danaher, 2016）不僅是理論性關切——它對學生申訴程序、法律責任及機構評鑑有實務意涵。

迭代推理導致的不可預測行為（**Unpredictable behavior through iterative reasoning**）。代理式 AI 系統以迭代方式推理，將複雜的評量任務分解為子任務，依序執行，並根據中間結果調整其方法。雖然此迭代能力使精密的評量策略成為可能，但它也引入了將代理式系統與規則導向替代方案區分開來的根本不可預測性。規則導向的早期預警系統產生確定性輸出：給定相同輸入，將始終產生相同警報。相反地，代理式 AI 系統可能在不同場合追求不同的推理路徑，產生難以預測、重現或解釋的評量結果。此隨機性（stochasticity）對作為評量實務基礎的一致性與公平性原則構成挑戰（Mislevy et al., 2003）。

持久記憶與偏誤複合（**Persistent memory and compounding bias**）。代理式 AI 用於學生成功衡量之架構的一個定義性特徵，是其維護縱貫式學生檔案的能力——跨課程、學期及學年累積資料以建構全面的能力軌跡。雖然此持久性使寶貴的縱貫分析成為可能，但它也創造了早期評量錯誤或偏誤判斷可隨時間複合的機制。一名在一年級某基礎能力被錯誤評估為不足的學生，可能發現此一初始錯誤形塑了所有後續的 AI 驅動評量，創造出低估的自我強化循環。此「偏誤複合」（bias compounding）風險有別於傳統演算系統中所記錄的靜態偏誤（Baker & Hawn, 2022），代表對公平評量的一種特別有害的威脅。

多代理協調失敗（**Multi-agent coordination failures**）。先進的代理式架構構想多個專業化 AI 代理協同完成評量任務——一個代理監測參與度，另一個評估書面作業，第三個追蹤能力進程。當這些代理成功協調時，結果是一個全面的評量生態系統。然而當協調失敗

時，錯誤責任被擴散至無法追溯的程度。若學生收到不正確的能力評量，判定哪個代理貢獻了該錯誤——以及如何更正——需要對可能不以可理解形式存在的多代理互動日誌進行鑑識分析。

目標錯位與指標遊戲化（**Goal misalignment and metric gaming**）。代理式 AI 最為隱伏的風險，或許是為「學生成功指標」而優化的系統可能學會優化指標本身，而非其旨在衡量的底層學習。此現象在運算脈絡中有時被稱為古德哈特定律（Goodhart's Law：「當衡量標準成為目標，它便不再是好的衡量標準」），當優化者是一個具備策略行為能力的自主代理時，便產生新的維度。一個因改善留存率而獲得獎勵的代理式 AI 系統，例如，可能學會降低評量門檻而非改善學習支持——一種改善指標但同時削弱其所聲稱服務之教育使命的策略。

6.4 代理式 **AI** 評量治理架構

因應前述所辨識的風險，需要一個同時在國家、機構及技術層級運作的治理架構。援引台灣現有的法律基礎建設、最近頒布的人工智慧基本法及國際最佳實務，本節提出三層治理架構（three-tier governance architecture）。

第一層：國家治理（***National Governance***）（教育部與 ***HEEACT***）

在國家層級，治理應錨定於台灣的人工智慧基本法，該法闡述七大指導原則：人類自主、隱私保護、透明、公平與不歧視、安全與保障、問責及永續。這些原則提供規範性基礎，但需要在教育評量的特定脈絡中加以操作化。

教育部應與 HEEACT 合作，制定「AI 評量標準」，將該法之原則轉化為對高等教育機構的具體要求。這些標準應強制要求 (a) 將所有用於重大評量決定的 AI 系統作為機構評鑑文件的一部分加以登錄，(b) 每年進行公平性影響評估，依人口統計特性——包括原住民身分、新住民背景、社經地位及身心障礙——分群分析 AI 系統績效，以及 (c) 最低透明度要求，明定學生在 AI 系統參與評量決定時必須被告知，且必須能夠獲得關於這些決定如何達成的解釋。HEEACT 作為台灣國家品質保證機構的既定角色，以及其 INQAAHE 正式會員地位，使其得以兼具國內公信力與國際接軌地制定這些標準。

第二層：機構治理（***Institutional Governance***）

在機構層級，每所部署代理式 AI 於評量中的大學應設立 AI 評量倫理委員會，以治理人類受試者研究的機構審查委員會（IRBs）為模型。此委員會應包括教師代表、學生代表、資料科學專家及外部倫理諮詢者。其職權應涵蓋 AI 評量系統的部署前審查、系統績效與公平性成果的持續監測，以及學生對 AI 驅動評量決定之申訴的裁決。

機構應進一步採行「學生資料權利宣言」（Student Data Bill of Rights），保障五項核心權利：(a) 知情同意——學生必須對 AI 驅動的評量表示肯認式同意，且拒絕者有真正的替代方案可用；(b) 存取——學生必須能夠查看所有關於自身的蒐集資料及由此衍生的能力評量；(c) 更正——學生必須能夠質疑並更正不準確的資料或評量；(d) 刪除——畢業或退學時，學生必須能夠請求刪除其行為及評量資料；(e) 可攜性——學生必須能夠以可互通的格式匯出其能力紀錄。台灣的個人資料保護法為這些權利提供法律基礎，但機構政策必須在教育脈絡中將該法的一般性條文加以領域特定的詮釋與操作化。

教師 AI 素養不應被視為可選的專業發展機會，而應作為任何其課程納入 AI 驅動評量之教師的先決條件。素養計畫不僅應涵蓋 AI 系統的技術操作，還應涵蓋其倫理意涵、局限性，以及在必要時詮釋與推翻 AI 建議所需的批判性判斷。此途徑將 AI 定位為增強專業能力的工具，而非取代專業能力的工具，直接回應倫理分析中所辨識的去專業化風險。

第三層：技術治理（***Technical Governance***）

技術治理要求應確保系統本身依循倫理原則而設計與運作。四項要求是不可或缺的。

第一，演算法透明度（algorithmic transparency）。部署代理式 AI 的機構必須能夠提供評量決定如何達成的有意義解釋。這未必需要完全的模型可解釋性（model interpretability）——對複雜的代理式系統而言可能在技術上不可行——但確實需要 Floridi et al. (2018) 所稱的「可解說性」（explicability）：提供關於為何做出特定決定的可理解說明的能力，即使完整的運算過程無法被完全透明化。

第二，偏誤測試方案（bias testing protocols）。AI 評量系統必須在部署前及持續運作基礎上接受偏誤測試。部署前測試應使用合成資料與歷史資料，評估系統在各人口群體間的表現。部署後監測應持續追蹤結果差異，並在差異超過預定門檻時觸發自動審查。Chinta et al. (2024) 所提出的 FairAIED 架構提供了技術上嚴謹的方法論用於此類測試，可改造適用

於台灣的人口統計脈絡。

第三，資料最小化（data minimization）。與個人資料保護法及人工智慧基本法之隱私原則一致，代理式 AI 系統應僅蒐集其評量功能所需的最少資料。必須抵抗因技術上可行便蒐集全面行為資料的誘惑，轉而以明確闡述之評量目的為指引進行有原則的資料蒐集。

第四，互通性標準（interoperability standards）。為保護學生的資料可攜性並防止供應商鎖定，AI 評量系統應遵循既定的學習資料標準，包括 xAPI 及 cmi5 用於學習紀錄，以及 Open Badges 用於能力認證。互通性不僅是技術上的便利——它是確保學生對其教育紀錄保有有意義控制的倫理要求。

本節所呈現的倫理分析與治理架構揭示，代理式 AI 在學生成功衡量中的部署不僅是一項技術性事業，更是一項根本的規範性事業。風險是真實的，證據基礎是不完整的，治理基礎建設也是初生的。然而台灣擁有獨特的制度優勢——HEEACT 的成熟品質保證生態系統、最近頒布的人工智慧基本法、強健的資料保護法律傳統，以及校際合作的文化——使其得以發展可作為區域典範的治理實務。問題不在於是否要在評量中接觸代理式 AI，而在於如何以尊重學生自主（autonomy）、教師專業及定義公正高等教育體系之公平承諾的方式來進行。下一節綜合所有研究問題的發現，並思考其對政策、實務與未來研究的意涵。

7. 討論與未來方向

7.1 整合 **ADAPT** 框架：關鍵洞見

本文探討代理式人工智慧（agentic AI）能否以及如何轉變台灣高等教育的學生學習成效衡量方式。第二節檢視代理式 AI 的技術能力，第三節盤點現行衡量典範的結構性限制，第四節提出 ADAPT 框架作為整合分析工具，第五節評估政策情境與評鑑建議，第六節進行倫理分析。在各環節分析完成後，本節綜合檢視這些環節作為整體所揭示的意涵。

最關鍵的發現在於技術能力與結構性需求之間的匯聚。第三節辨識出六項限制：時間性、粒度、模態、能動性、素養捕捉，以及間接衡量的主導地位。這些限制並非隨意的缺陷，而是為定期審查、文件佐證與整體性判斷所設計之評量架構的可預見結果。第二節辨識的代理式 AI 六項能力——自主規劃、動態適應、工具使用、多步驟推理、持久記憶與多代理協作——與這些限制形成高度直接的對應。持久記憶以持續而非定期的證據蒐集回應時

間性限制。個別學習者建模在學生層級而非學程層級追蹤學習軌跡，回應粒度限制。多模態證據整合在學習產出之外同時捕捉學習歷程，回應模態限制。自主觀察產生非完全由機構策展的證據，回應能動性限制。多步驟推理評估批判思考、創意問題解決、倫理推理等抗拒簡化為可計量產出的複雜素養，回應素養捕捉限制。而從問卷式到表現式證據的轉變，則回應間接衡量居主導地位的問題。

此一匯聚正是孔恩式框架所預測的現象。典範轉移之所以發生，並非因為新理論在抽象層面上更為優越，而是因為它能解決既有典範無法處理的特定異例 (Kuhn, 1962/2012)。第四節提出的 ADAPT 框架將此匯聚加以操作化，描繪出從異例辨識（「D」——診斷性盤點）、經由評量重構（第二個「A」——評量再概念化）、到政策實施（「P」——政策路徑）及倫理保障（「T」——信任與倫理保障）的路徑。該框架的分析效用在於其整合性：它不將技術、政策與倫理視為各自獨立的對話，而是將其視為單一轉變中相互依存的面向。

第二個關鍵洞見涉及第四節辨識的七個典範轉移面向。這些面向涵蓋時間性（從定期到持續）、粒度（從總體到個別）、能動性（從機構自陳到 AI 觀察與學生共創）、回饋延遲（從回溯到即時）、證據類型（從文件與問卷到多模態學習軌跡）、評量目的（從績效責任到品質改善），以及評量者身分（從純人工到人機協作）。這些面向描述的不僅是技術升級，更是一種根本不同的學習衡量認識論。現行典範的核心問題是：「該機構是否透過其自行策展的文件證據，證明其已建立學習成效評量機制？」新興典範的核心問題則是：「學生學習中實際發生了什麼？持續性、多來源、多模態的證據流是否已捕捉此一實況，且未被任何單一利害關係人所掌控？」這是從衡量機構合規到衡量教育實況的轉移，對評鑑實務具有深遠意涵。

第三個洞見涉及 ADAPT 框架超越台灣脈絡的可移轉性。儘管本文將該框架應用於代理式 AI 與 HEEACT 評鑑體系的交會處，其五層分析結構——界定技術、診斷限制、重構評量、評估政策路徑、進行倫理治理——適用於任何面臨相同技術衝擊的品質保證體系。東南亞 AQAN 成員機構、印度 NAAC 及阿聯酋 CAA 等品質保證機構面臨結構上類似的挑戰：定期評鑑週期、文件式證據，以及持續性品質監控能力的不足。ADAPT 框架為這些機構提供結構化的分析路徑，檢視代理式 AI 如何與其評量架構交互作用，而非提供一體適用的解決方案。未來的比較研究可將該框架應用於不同國家脈絡，辨識典範轉移中哪些

面向具有普遍性，哪些具有脈絡依存性。

7.2 本文未主張之事項

學術誠信要求明確界定本分析所主張與未主張之事項。以下四項界限尤為重要。

第一，本文並未主張代理式 AI 必然會轉變高等教育評量。本文所理論化的典範轉移是一種可能性，而非預言。它取決於機構準備度、教師接受度、監管意志、基礎設施投資，以及技術成熟速度等偶然因素，其中任何一項都可能使轉變停滯或改變方向。第五節討論的南韓案例顯示，即便政府積極承諾 AI 部署，仍可能因利害關係人的抵制而逆轉。本文力求避免技術決定論——即假設技術能力會自動轉化為機構採納。

第二，本文並未主張台灣現行評量典範毫無價值。HEEACT 評鑑框架歷經三個週期、逾十五年的發展與精進，是成熟且獲國際認可的品質保證體系。該體系在台灣高等教育部門中培育了自我評鑑、持續改善與證據本位決策的文化 (Lin et al., 2021)。本文的論點並非揚棄此框架，而是使其進化。其對品質、公平與績效責任的根本承諾，可透過運用原始設計者在當時無法預見之技術能力的評量架構，獲得更充分的實現。

第三，本文並未主張 AI 應取代教育評量中的人類判斷。第六節的倫理分析及其治理框架建立在相反的信念之上：教師的專業知能、評鑑委員的集體審議、源自與教育社群持續互動的脈絡理解，這些人類判斷不可化約也不可替代。代理式 AI 可透過提供更豐富、更即時、更細緻的證據來增強此判斷，但它無法也不應取代唯有人類方能提供的教育評鑑之規範性與關係性面向。

第四，本文並未提供代理式 AI 改善學習成效的實證證據。作為理論與政策分析論文，本文建構概念框架、評估政策情境並提出治理結構。這些框架的實證驗證——透過縱貫性實施研究、隨機對照試驗及準實驗設計——屬後續研究之工作，而非本文之任務。

7.3 分析之限制

本分析受若干限制所制約，以下逐一說明。

最根本的限制在於本文的理論性格。本文主張代理式 AI 能回應現行典範的結構性限制、ADAPT 框架能準確描繪典範轉移面向、所建議的政策路徑能在創新與審慎之間取得平衡。這些主張奠基於理論推理與類比論證，而非實證觀察。在代理式 AI 評量系統於台灣高等教育機構中實際實施並經嚴謹評估之前，這些主張仍屬暫時性的。正如 UNESCO

(2023) 所警示，AI 在教育中的證據基礎「仍在形成中，且就許多應用而言大致是不確定的」。此一警示對於較生成式 AI 更新且研究更少的代理式 AI，更具適用力。

第二，本分析具有台灣特定性。ADAPT 框架參照 HEEACT 評鑑結構、教育部政策工具、台灣《人工智慧基本法》，以及台灣高等教育體系的人口、經濟與制度條件而發展。儘管第 7.1 節論證了該框架更廣泛的適用性，此適用性在其他國家脈絡中經過實際檢驗之前仍屬推測。各國品質保證體系在監管權限、文化脈絡、機構多樣性與技術基礎設施方面差異顯著，適用於台灣的方案未必能直接移轉至其他場域。

第三，技術圖景的演進速度對任何以傳統學術格式發表的分析構成挑戰。今日仍屬推測性的代理式 AI 能力，在本文發表時可能已成常態；目前看似前景可期的能力，也可能遭遇尚未顯現的技術障礙。第二節提出的四層分類法儘管基於 2026 年初可得之最佳證據，隨著技術成熟仍可能需要修訂。

第四，本文缺乏第一手利害關係人資料。教師、學生、機構行政人員、雇主及評鑑委員——即最直接受本文所理論化之典範轉移影響的群體——其觀點僅透過二手文獻呈現。全面的政策分析理想上應納入對這些利害關係人的結構化諮詢、問卷調查或訪談。本文承認此一缺口，未來研究須加以填補。

第五，本分析未深入探討大規模實施代理式 AI 評量系統的技術可行性。運算基礎設施、資料架構、系統互操作性及維護成本等問題雖經認定為關鍵，處理層次較為概括，未能充分反映其複雜性。第五節提出的分階段實施路徑預設這些技術挑戰可以處理，此一預設須透過試辦加以驗證。

7.4 未來研究議程

上述限制從反面界定了一個範疇可觀且具急迫性的研究議程。以下六項優先事項值得特別關注。

優先事項一：代理式 **AI** 評量的縱貫性實證研究。最迫切的研究需求是來自實際實施的實證證據。第五節建議的第一階段結構化試辦應伴隨嚴謹、獨立的評估研究，採用混合方法設計：以量化方法分析學習成效的效度與信度、以質性方法探究教師與學生經驗，並以準實驗方法比較傳統評量。這些研究應橫跨多個學期，以捕捉本文理論論證核心的縱貫動態。從快照到軌跡的轉移無法透過短期試辦驗證。

優先事項二：弱勢群體的公平影響研究。第六節的倫理分析辨識出演算法偏差與數位落差擴大為高可能性、高嚴重性的風險。這些風險最沉重地落在特定群體身上：原住民學生、新住民子女、身心障礙學生，以及社經弱勢背景的學生。專門的公平影響研究應檢視代理式 AI 評量系統是否對這些群體產生差異性結果。若是，則應進一步探究何種設計修改與治理干預措施能減緩此類差距。Baker and Hawn (2022) 及 Gandara et al. (2024) 關於美國教育 AI 系統中演算法偏差的研究，提供了可調適至台灣人口脈絡的方法論模型。

優先事項三：教師經驗與接受度研究。教師是任何評量轉變的關鍵樞紐。教師對代理式 AI 工具的接受、抗拒、調適與創造性挪用，將決定本文所理論化的典範轉移究竟能成為制度現實，抑或僅止於理論願景。運用科技接受模式（Technology Acceptance Model, TAM）、延伸納入 AI 素養的 TPACK 框架，以及 AI 增強評量脈絡下教師專業認同的質性研究，將為政策設計提供關鍵證據。

優先事項四：**AI** 增強品保框架的跨國比較研究。隨著各國品質保證機構面對 AI 對評鑑的意涵，比較研究可辨識哪些挑戰與解方具有普遍性，哪些具有脈絡特定性。針對 HEEACT（台灣）、TEQSA（澳洲）、QAA（英國）、CHEA（美國），以及新加坡和南韓新興 AI 增強品保框架所採取之方法進行系統性比較，將豐富政策學習與制度調適的證據基礎。

優先事項五：透過個案研究驗證 **ADAPT** 框架。第四節提出的 ADAPT 框架是需要實證驗證的理論建構。將該框架應用於特定機構實施案例的個案研究——追蹤從代理架構分析、經由診斷性盤點、評量重構、政策路徑選擇，到信任保障部署的進程——將檢驗該框架的分析效用，並辨識需要精進之處。

優先事項六：**AI** 評量互操作性的技術標準。第六節的治理框架辨識出互操作性既是技術需求，也是倫理要求。針對現有學習資料標準——xAPI、cmi5 及 Open Badges——在代理式 AI 評量情境中的應用進行研究，以及在現有標準不敷使用時發展新標準，對於確保 AI 生成之學習證據的可攜性、可稽核性與跨機構互操作性至關重要。此類標準亦將促進系統層級品質保證所需的跨機構資料彙整。

8. 結論

本文始於一個悖論：一個以培育學生因應快速變遷世界為志的高等教育體系，卻以僅能漸進改變的機制來衡量學習成效。本文終於一個命題：代理式 AI 能力與台灣評量架構中累

積的結構性限制之匯聚，為典範轉移創造了條件。這不是對既有方法的漸進改良，而是對「衡量學生學習意味著什麼」的根本重構。

評量領域中典範層級變革的技術能力正在快速浮現，為台灣品質保證體系帶來必須積極面對的機會與風險。代理式 AI 系統在全球教育場域中的部署可能性日益增高，驅動力來自尋求競爭優勢的機構、尋求市場份額的科技公司，以及尋求個人化支持的學生。台灣品質保證社群面臨的核心問題是：此一交會將如何管理？是透過審慎治理還是臨時因應？是透過前瞻性的框架設計還是事後的修補拼湊？是透過以公平為核心的政策，還是透過深化既有機構落差的市場驅動擴散？

本分析的三項關鍵發現，在台灣接近此一決策點之際值得特別強調。

第一，代理式 AI 為過去不可能實現的學習衡量形式創造了技術前提。跨課程與跨學期的持續素養追蹤、校準至個別學習者的個人化評量策略、在學習產出之外同時捕捉學習歷程的多模態證據整合，以及揭示素養如何浮現與鞏固的縱貫性發展軌跡——這些並非對既有工具的漸進改良，而是質性上全新的衡量模態。它們精準回應了現行典範無法從其自身邏輯內部解決的結構性限制。過去以六年回溯性快照所衡量之事物，原則上可以作為持續演進的軌跡來衡量。然而，實現此潛力需要本文所倡議的審慎、證據本位的實施方式。

第二，此一轉移需要的不僅是技術，更需要治理。第六節的倫理分析顯示，代理式 AI 在評量中的風險——演算法偏差、監控常態化、教師去專業化、課責稀釋，以及持久記憶偏差的複利效應——並非附帶的副作用，而是該技術自主性、適應性與持久性的內在特徵。本文提出的三層治理框架——以《人工智慧基本法》為錨點的國家標準、以研究倫理委員會為模型的機構倫理委員會，以及要求透明性、偏差檢測、資料最小化與互操作性的技術治理——並非 AI 部署的選配，而是負責任部署的先決條件。缺乏治理的技術不是創新，而是失職。

第三，台灣擁有足以引領的制度優勢。HEEACT 的國際地位——完全符合 INQAAHE 標準、獲 CHEA 認可、積極參與 APQN——提供了一個平台，使台灣的 AI 增強品質保證方法得以影響區域及全球實踐。《人工智慧基本法》提供了許多同儕國家所缺乏的法律基礎。歷經三個評鑑週期所培育的成熟品質保證文化，涵蓋逾 140 所機構對證據、評鑑與改善的共同理解，為 AI 增強評量提供了可資建構的制度基底。台灣的少子女化危機雖深具挑戰

性，卻創造了清晰的目標感：在機構存續取決於教育價值展現的體系中，發展更有效、更即時、更具行動力的學習成效衡量，不是理論問題，而是攸關存亡的現實。

第四週期校務評鑑代表決定性時刻。第五節提出的具體建議——修訂核心指標第三項標準之描述語、於自我評鑑報告中新增「AI 於評量之應用」附錄、調整實地訪評程序（含 AI 系統稽核檢核表）、以自我評鑑機構為早期採用者，以及三組件式評鑑委員增能方案——均可在第四週期設計窗口內實施，無需立法行動或高不可攀的基礎設施投資。這些建議代表的是框架演進路徑：企圖心足以使台灣在亞太地區 AI 增強品質保證領域居於前沿，審慎度則足以避免南韓較為激進的方式所遭遇的利害關係人反彈與公平問題。

據此，本文提出具體的行動呼籲：在 2026 至 2028 年的窗口期內啟動結構化試辦，以審慎態度與嚴謹的公平監測推進，透過國際標竿學習與專業化培訓建立評鑑委員的能力，並將成果嵌入第四週期框架之中。此框架應正式認可 AI 生成的學習證據，同時保存人類判斷、機構自主性與改善導向——這些正是台灣品質保證傳統的核心價值。完美的證據永遠不會出現，但現有的證據基礎結合本文的理論分析，已足以支持結構化試辦的正當性。本文倡議的不是全面採用，而是在能偵測並修正錯誤的治理框架內進行有紀律的實驗。在缺乏治理的情況下貿然部署固然風險重大，在同儕體系持續前進的同時無限期延遲行動同樣代價高昂。審慎的路線是在台灣法律與制度架構所能支撐的治理結構下，把握當前的政策窗口採取行動。

本文標題援引了一個隱喻：從快照到軌跡的移動。快照凝固一個瞬間，呈現某人在某一時間點上所處的位置。軌跡描繪一條路徑，揭示某人曾到過哪裡、正往何處去，以及其移動方式如何隨時間改變。長久以來，台灣的品質保證體系以快照的方式衡量學生學習——定期的、靜態的、回顧性的。能夠實現軌跡式衡量的技術正在浮現——持續的、動態的、前瞻性的。此一願景並非必然的結果，它取決於機構的承諾、充足的資源、有效的治理，以及教師與學生參與新評量模態的意願。本文提出的治理框架、倫理保障與政策路徑，旨在確保此一轉變若發生時，能服務台灣高等教育體系中的每一位學生——不僅是資源充裕機構中的學生，不僅是學習行為符合演算法預期的學生，而是每一位教育軌跡值得被看見、被理解、被支持的學生。從快照到軌跡的轉移，歸根究柢，是從衡量機構到理解學習者的轉移。這是一個值得追求的典範——以嚴謹、以謙遜，以及對證據的堅定關注。

References

Agent4EDU. (2024). Advancing AI for education with agentic workflows. In *Proceedings of the ACM International Conference on AI in Education (ICAIE '24)*. ACM.

Association of American Colleges and Universities. (2018). *Fulfilling the American dream: Liberal education and the future of work*. AAC&U.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Banihashem, S. K., Gasevic, D., Noroozi, O., Jarodzka, H., Joosten-ten Brinke, D., & Drachsler, H. (2025). Optimizing formative assessment with learning analytics: A systematic review. *Review of Educational Research*, *95*(2), 215–258.

Bardach, E., & Patashnik, E. M. (2019). *A practical guide for policy analysis: The eightfold path to more effective problem solving* (6th ed.). CQ Press.

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160–1173. https://doi.org/10.1111/bjet.1

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

Biggs, J. (1999). *Teaching for quality learning at university*. Society for Research into Higher Education & Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Le Quy, T., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:2407.18745*.

Coates, H., & Zlatkin-Troitschanskaia, O. (2019). The governance, policy and strategy of learning outcomes assessment in higher education. *Higher Education Policy*, *32*, 507–512. https://doi.org/10.1057/s41307-019-00161-1

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change.

Educational Researcher, *32*(6), 3–12. https://doi.org/10.3102/0013189X032006003

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, *29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The rise of agentic AI. *Future Internet (MDPI)*, *17*(9), 404.

ENQA. (2015). *Standards and guidelines for quality assurance in the European Higher Education Area (ESG)*. European Association for Quality Assurance in Higher Education.

Eckstein, H. (1992). *Regarding politics: Essays on political theory, stability, and change*. University of California Press.

Ewell, P. T. (2009). Assessment, accountability, and improvement: Revisiting the tension (NILOA Occasional Paper No. 1). National Institute for Learning Outcomes Assessment.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gandara, D., Anahideh, H., Ison, M. P., & Picchiarini, L. (2024). Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *AERA Open*, *10*(1), 1–15. https://doi.org/10.1177/23328584241258741

Gartner. (2025). *Top strategic technology trends for 2025: Agentic AI*. Gartner, Inc.

Goffman, E. (1959). *The presentation of self in everyday life*. Anchor Books.

HEEACT. (2023a). *Third cycle of institutional accreditation handbook (2023–2025)*. Higher Education Evaluation and Accreditation Council of Taiwan.

HEEACT. (2024). *Program accreditation handbook (2024 edition)*. Higher Education Evaluation and Accreditation Council of Taiwan.

Hou, A. Y. C., Morse, R., & Chiang, C. L. (2012). An analysis of mobility in global rankings: Making institutional strategic plans and positioning for building world-class universities. *Higher Education Research & Development*, *31*(6), 841–857.

IMDA. (2020). *Model AI governance framework* (2nd ed.). Infocomm Media Development Authority, Singapore.

Inside Higher Ed. (2026, January). AI agent "Einstein" passes university courses autonomously: What it means for assessment design. *Inside Higher Ed*.

Kasparov, G. (2017). *Deep thinking: Where machine intelligence ends and human creativity begins*. PublicAffairs.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms active learning. *Scientific Reports*, *15*, 17458. https://doi.org/10.1038/s41598-025-97652-6

Kuhn, T. S. (2012). *The structure of scientific revolutions* (4th ed.). University of Chicago Press. (Original work published 1962)

Legislative Yuan. (2025). *Artificial Intelligence Basic Act* (人工智慧基本法). Republic of China (Taiwan).

Lin, A. S. R., Hou, A. Y. C., Chan, S. J., & Chiang, T. L. (2021). Quality assurance in Taiwan higher education: Regulation, model shift, and future prospect. In A. Y. C. Hou, T. L. Chiang, & S. J. Chan (Eds.), *Higher Education in Taiwan: Global, political and social challenges and future trends* (pp. 65–81). Springer. https://doi.org/10.1007/978-981-15-4554-2_4

Masterman, M. (1970). The nature of a paradigm. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 59-89). Cambridge University Press.

Masterman, T., Besen, S., Sawtell, M., & Chao, A. (2024). The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*.

Ministry of Education. (2024). *Education statistical indicators at a glance 2024*. Ministry of Education, Republic of China (Taiwan).

Ministry of Education. (2025). *2025 Education White Paper*. Ministry of Education, Republic of China (Taiwan).

Ministry of Education Singapore. (2023). *EdTech Masterplan 2030: Technology-transformed learning to prepare students for the future*. Ministry of Education, Singapore.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design (ETS Research Report No. RR-03-16). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Ng, A. (2024, March 13). Agentic design patterns. *The Batch* (DeepLearning.AI Newsletter).

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509.

OECD. (2023). *OECD digital education outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing. https://doi.org/10.1787/c74f03de-en

OpenAI. (2025, July). New tools for understanding AI and learning outcomes. `https://openai.com/index/understanding-ai-and-learning-outcomes/`

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, Article 100020. https://doi.org/10.1016/j.caeai.2021.100020

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

QAA. (2023). *Artificial intelligence: Guidance for UK higher education providers*. Quality Assurance Agency for Higher Education.

Rest of World. (2025, June). South Korea scales back AI textbook rollout after parental backlash. *Rest of World*.

Ritzer, G. (1975). Sociology: A multiple paradigm science. *The American Sociologist*, *10*(3), 156-167.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Sharma, Y. (2024, September). Taiwan faces wave of university closures. *University World News*.

Shavelson, R. J. (2010). Measuring college learning responsibly: Accountability in a new era. *Stanford University Press*.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1509–1528. https://doi.org/10.1177/0002764213479366

Spencer, L. M., & Spencer, S. M. (1993). *Competence at work: Models for superior performance*. John Wiley & Sons.

Yan, L. (2025). From passive tool to socio-cognitive teammate: Reconceptualizing AI's role in learning through the lens of agentic cognition. *arXiv preprint arXiv:2508.14825*.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., … & Gasevic, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075

Taiwan News. (2024, August). Seven Taiwan universities shut down amid enrollment crisis. *Taiwan News*.

Tam, M. (2001). Measuring quality and performance in higher education. *Quality in Higher Education*, *7*(1), 47–54. https://doi.org/10.1080/13538320120045076

Arunkumar, V., Gangadharan, G. R., & Buyya, R. (2026). Agentic AI: Architectures, taxonomies, and evaluation of LLM agents. *arXiv preprint arXiv:2601.12560*.

Temper, M., Tjoa, A. M., & David, K. (2025). Higher Education Act for AI (HEAT-AI): A framework to regulate the usage of AI in higher education institutions. *Frontiers in Education*, *10*, Article 1505370. https://doi.org/10.3389/feduc.2025.1505370

TEQSA. (2024). *Artificial intelligence in higher education: Guidance note for providers*. Tertiary

Education Quality and Standards Agency, Australian Government.

UNESCO. (2023). *Guidance for generative AI in education and research*. United Nations Educational, Scientific and Cultural Organization.

UNESCO. (2025, October 27). What's worth measuring? The future of assessment in the AI age. https://www.unesco.org/en/articles/whats-worth-measuring-future-assessment-ai-age

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Zhong, L., & Zhao, X. (2025). Education paradigm shifts in the age of AI: A spatiotemporal analysis of learning. *ECNU Review of Education*, *8*(2), 319–342. https://doi.org/10.1177/20965311251315204

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.